



**An Investigation into the Effects of Topic and
Background Knowledge of Topic on Second Language
Speaking Performance Assessment in Language
Proficiency Interviews**

Nahal Khabbazzashi
Worcester College

Thesis submitted to the University of Oxford
for the degree of Doctor of Philosophy

Department of Education, University of Oxford

January 2013 (Michaelmas Term)

Abstract

This study explores, from a test validity perspective, the extent to which the two variables of topic and background knowledge of topic have an effect on spoken performance in language proficiency interviews. It is argued that in assessment contexts where topics are randomly assigned to test takers, it is necessary to demonstrate that topics of tasks and the level of background knowledge that test takers brings to these topics do not exert an undue influence on test results. Otherwise, a validity threat may be introduced to the test.

Data were collected from 82 Farsi speakers of English who performed on ten different topics, across three task types. Participants' background knowledge of topics was elicited using self-report questionnaires while C-tests were used as a measure of general English language proficiency. Four raters assigned scores to spoken performances using rating scales. Semi-structured interviews were carried out with raters, upon completion of the rating process. A mixed-methods strategy of inquiry was adopted where findings from the quantitative analyses of score data (using Multi-Faceted Rasch Measurement, multiple regression and descriptive statistics) were synthesised with the results of the qualitative analyses of rater interviews and test takers' content of speech in addressing the foci of the study.

The study's main findings showed that the topics used in the study exhibited difficulty measures which were statistically distinct i.e. topics, within a given task type, could not be considered parallel. However, the size of the differences in topic difficulties was too small to have a large practical effect on scores. Participants' different levels of background knowledge were shown to have a consistent, systematic and statistically significant effect on performance with low levels of background knowledge posing the highest level of challenge for test takers and vice versa. Nevertheless, these statistically significant differences in background knowledge levels failed to translate into practically significant differences, as the size of the differences were too small to have a large impact on performance scores. Results indicated that, compared to general language proficiency which accounted for approximately 60% of the variance in spoken performance scores, background knowledge only explained about 1-3% of the variance.

Qualitative analyses of data suggested lack of background knowledge to be associated with topic abandonment, disengagement from topic-related questions, and fewer opportunities for test takers to elaborate on topics. It was also associated with negative affective influence on test takers, particularly lower proficiency individuals. Taken together, the findings have theoretical, methodological and practical implications for second language speaking performance assessment.

*Dedicated with love and gratitude to my Momzooki and to the memory of my grandfather,
Baba Madani, and the loved ones who are no longer with me.*

Acknowledgments

I wish to thank the wonderful people who have helped and supported me throughout this unforgettable DPhil journey and without whom this thesis would not have seen the light of day or made its way to the depths of the Bodleian library.

My deepest respect and most profound gratitude go to my supervisor, Dr. Bob Vanderplank, for his guidance, support and encouragement and for always believing in me. I am incredibly blessed to have had such an amazing mentor and will remain forever thankful for all his kindness.

I am indebted to Professor David Andrich, a truly inspiring academic, for teaching me everything I know about Rasch analysis (and the Black Swan Theory!) and for helping me with my data analyses. I am grateful to my co-supervisor, Dr. Catherine Walter, for her helpful advice and detailed feedback on several chapters of the thesis. I owe a special thank you to my friend and colleague, Dr. Gad Lim, for not only reading and commenting on the entire first draft of the thesis but for supplying me with much needed chocolate during stressful times. My heartfelt gratitude also goes to Dr. Evelina Galaczi for her kindness and continuous support ever since I moved to Cambridge and for generously sharing her expertise in speaking assessment with me.

I would like to thank Professor Ingrid Lunt for my doctoral training in the first year of my studies. I am grateful to Dr. Victoria Murphy, Professor Ernesto Macaro and Professor David Andrich for their helpful comments and critical feedback at the transfer and confirmation stages of the DPhil. I would also like to thank Professor Barry O’Sullivan and Professor Ernesto Macaro for examining my final thesis, for their invaluable insights and for a truly memorable viva experience.

I am indebted to all the participants in this project for their time and collaboration. This thesis would not have been completed without them. I would like to thank Ms. Mandana Arfa and Dorothea for their help in my data collection and Ron, Margaret, Jess and Sara for rating all the data. I am grateful to the members of the applied linguistics research group, professors, colleagues, library and admin staff, and fellow DPhilers and StANers at the Department of Education, University of Oxford as well as all the wonderful members and staff of Worcester College for all their support during my studies in Oxford. Special thanks go to Ms. Cath Fraser, our graduate officer, for all her help since my first day at college.

I am grateful for the financial support of the TOEFL Grants and Awards Committee in offering me the TOEFL Small Grants for Doctoral Research in Second or Foreign Language Assessment and to the Department of Education, University of Oxford for the 2nd Year Research Bursary.

They say a DPhil is a lonely journey. I beg to differ. I have been lucky enough to have family and friends who have been there through all the good times and the bad, who have shared with me the drama, pain and laughter that goes hand in hand with doing a DPhil. Thank you to all my dudes at Oxland and Camtown for the wonderful memories over the past couple of years, for late teas at Franks, for drinks at the Duke, for knowing when caffeine levels are dangerously low, for the inside jokes, for writing the N-N agreement, for ensuring that I boogie my way to calmness, for geeking out at conferences, for making sure that I eat, for helping me with technology, for indulging me in my movie obsessions, for putting up with the rabbit in Alice in Wonderland, for introducing me to momos, for letting me go on and on about my thesis, for staying up for the Oscars, for the randomness and most of all, for keeping me sane! Thank you Dani, Boo, Tania, Natmouchan, Lady J, Ron, Megan, Ms. Annie B, Pegah, Muneera, Matthias and Kasiunka. A special thank you also goes to the Combibos family for giving me a second home (and the best coffee!) in Oxford.

I owe a debt of gratitude to my family and friends back home for their loving support and encouragement from thousands of miles away. They brightened my long and gloomy days at the library with their calls, messages, e-mails, gossip and news; they spoiled me endlessly every time I went home and gave me strength to carry on. Thank you Baba Nimz, Khaleh Noonzi, my little Joojoo, Maman Forough, Maman Nahid, Haleh, Dae, Feri, Meliss, Khaleh Faf, Amoo Kag, Bizoo, Mona, Sogoli, Amir, Negari, Tolangi, Khaleh Felor Zare & Najafzadeh, Amoo Anoosh, Khaleh Mina, Mandana joon, Dr. Parsi and Goli joon for being there and for cheering me to the finish line!

My mooshoo, Gunnar, deserves special thanks for his patience in dealing with stressed, frustrated, weird and confused versions of me; for enlightening me on the merits of RTFMing; for occasional roosting and most of all, for always being there.

Above all, I would like to thank my Momzooki and Babaee for their love and endless support and for going through the ups and downs of this journey with me, as if it were their own. Momzookam, you are the strongest and most amazing woman I know. Thank you for teaching me the value of perseverance and hard work, for never losing faith in me and for making me send in that application in the first place. Babaee, you are the gentle soul that brings calm and balance to my life. Thank you for teaching me the value of kindness and generosity and for reminding me of the importance of putting things into perspective. I would not have made it so far without the two of you. Thank you.

Table of Contents

Abstract	i
Acknowledgments	iii
Table of Contents	v
List of Tables	ix
List of Figures	xi
List of Appendices	xiii
Abbreviations and Acronyms	xiv
Chapter 1 Introduction	1
1.1 Variability and Spoken Test Performance	2
1.2 Speaking Task Characteristics: Topic	4
1.3 Test-taker Characteristics: Background Knowledge of Topic	7
1.4 Statement of the Problem	8
1.5 What is IELTS?	9
1.6 The IELTS Speaking Test	12
1.6.1 Test Format, Duration and Settings	12
1.6.2 Test Parts, Topic Sets and the Examiner Script	12
1.6.3 Test Construct, the Rating Scale and the Rating Process	14
1.6.4 The IELTS Speaking Test: a Critical Review	16
1.7 Rationale for the Study	20
1.8 Research Questions	22
1.9 Structure of the Thesis	24
Chapter 2 Literature Review	26
2.1 Topic and Background Knowledge of Topic: Impact on Performance	26
2.1.1 Topic as a Task Condition: Impact on Task Difficulty	26
2.1.2 Task Difficulty and Language Testing	31
2.1.3 Problematising Task Difficulty	34
2.1.4 Topic and Background Knowledge of Topic: Influence on L2 Performance	35
2.1.4.1 Reading	36
2.1.4.2 Listening	42
2.1.4.3 Writing	47
2.1.4.4 Speaking	51
2.1.4.5 Summary and Methodological Implications	57
2.1.5 Rater and Test-Taker Perspectives: Research from IELTS	59
2.2 The Rater Factor: Impact on Performance	62
2.3 Performance Assessment and Multi-Faceted Rasch Measurement	69
2.3.1 Approaches to Measurement: a Case of Different Paradigms?	70
2.3.2 What is Rasch Analysis?	71
2.4 Topic Validity and Spoken Performance Assessment	76
2.4.1 Validity	76
2.4.2 Validation	81
2.4.3 Weir's Socio-Cognitive Framework of Language Test Validation	82
2.4.4 Application of the Weir (2005) Framework to the Current Study	85

Chapter 3 Methodology	93
3.1 Approach to Research	94
3.2 Research Setting	95
3.3 Data Collection (Phase I)	96
3.3.1 Participants (Test Takers)	96
3.3.2 Instruments	96
3.3.2.1 Speaking Tasks/Speaking Test Forms	96
3.3.2.1.1 Number of Speaking Tasks/Parallel Forms	96
3.3.2.1.2 Speaking Task Selection	98
3.3.2.1.3 Establishing Task Equivalence	99
3.3.2.2 Background Knowledge Questionnaire	105
3.3.2.2.1 Assessing Background Knowledge of Topics	105
3.3.2.2.2 Background Knowledge Questionnaire	106
3.3.2.3 C-tests	108
3.3.2.3.1 Assessing General Language Proficiency	108
3.3.2.3.2 What is a C-test?	109
3.3.2.3.3 The Construct Validity of C-tests	109
3.3.2.3.4 C-test Design and Piloting	111
3.3.3 Procedures	111
3.3.3.1 Participant Selection Process	111
3.3.3.2 Data Collection Procedure	113
3.3.3.2.1 C-test Administration	113
3.3.3.2.2 Speaking Test Administration	113
3.3.3.2.3 Questionnaire Administration	115
3.3.3.3 Data Processing	115
3.3.3.3.1 Spoken Data	115
3.3.3.3.2 Background Knowledge Questionnaire Responses	116
3.3.3.3.3 C-test Responses	116
3.4 Data Collection (Phase II)	117
3.4.1 Rater Selection Criteria	117
3.4.2 Instruments	119
3.4.2.1 Rating Scales	119
3.4.2.2 Observation Checklist	121
3.4.2.3 Instructions/Score Sheets	122
3.4.3 Rating Procedures	122
3.4.3.1 Allocating Speaking Tasks to Raters	122
3.4.3.2 Rater Familiarisation Process	123
3.4.3.3 Further Clarifications and Batch Assignment	124
3.4.3.4 Post-rating Semi-structured Interviews	124
3.4.4 Data Processing	124
3.4.4.1 Speaking Scores Data Set	124
3.4.4.2 Interview Transcription	125
3.5 Quantitative Data Analysis	125
3.5.1 The Rasch Models	125
3.5.1.1 Background Knowledge Questionnaires	125
3.5.1.2 C-tests	126
3.5.1.3 Speaking Score Data	126
3.5.2 Multiple Regression Analysis	126
3.5.3 Frequency Statistics	127
3.6 Qualitative Data Analysis	127
3.7 Summary	127

Chapter 4 Preliminary Analyses	130
4.1 C-tests	130
4.1.1 C-test Pilot Study: Overview	130
4.1.2 C-test Pilot Study: Analysis using the Rasch model	131
4.1.3 C-test Pilot Study: Item Analysis	132
4.1.4 C-test Pilot Study: Super-item Analysis	138
4.1.4.1 Violation of the Assumption of Response Independence in C-tests	138
4.1.4.2 Results of C-test Super-Item Analysis	139
4.1.5 C-test Results (Main Study)	143
4.2 Background Knowledge Questionnaires	147
4.2.1 An Overview of the Questions, Scoring and Initial Observations	147
4.2.2 Measurement and Likert Scales	149
4.2.3 Background Knowledge Questionnaire Results	150
4.2.3.1 Summary Statistics	151
4.2.3.2 Individual Item Fit Statistics	152
4.2.3.3 Diagnosing Misfit: Thresholds and Category Functioning	153
4.2.3.4 Diagnosing Misfit: Differential Item Functioning (DIF)	156
4.3 Summary	162
Chapter 5 Findings	163
5.1 Introduction	163
5.2 MFRM Analysis: a Conceptual-Psychometric Framework	164
5.2.1 Applying the Framework: MFRM Analyses in the Current Study	167
5.3 Four-Facet MFRM Analysis: Results	167
5.3.1 Facet Map	168
5.3.2 The Examinee Facet	172
5.3.3 The Rater Facet	181
5.3.3.1 Rater Severity/Leniency and Consistency	182
5.3.3.2 Other Systematic Rater effects: Extremism, Central Tendency and Halo	187
5.3.4 The Criterion Facet	196
5.3.5 The Topic Facet	198
5.3.5.1 Examinee Ability and Topic Difficulty Distributions	199
5.3.5.2 Topic Measurement Report	201
5.3.5.2.1 FullA: Analysis with All Criteria	201
5.3.5.2.2 IELTSA: Analysis with IELTS Criteria	204
5.3.6 Rating Scales	207
5.3.7 Topic Influence on Performance Scores (Test Level)	214
5.3.7.1 FullA: Analysis with All Criteria	214
5.3.7.2 IELTSA: Analysis with IELTS Criteria	216
5.3.8 Topic Influence on Performance Scores (Task Level)	219
5.3.8.1 Task Type A: Topic Measurement Reports	219
5.3.8.2 Task Type B: Topic Measurement Reports	221
5.3.8.3 Task Type C: Topic Measurement Reports	223
5.3.8.4 Topic Influence on Scores at the Task Level (FullA)	225
5.3.8.5 Topic Influence on Scores at the Task Level (IELTSA)	227
5.3.9 Functions Analyses	230
5.3.10 Global Model Fit	236
5.4 Five-Facet MFRM Analysis	238
5.4.1 Facet Map	239
5.4.2 The Background Knowledge Facet	242
5.5 Multiple Regression Analysis: Background Knowledge as a Predictor Variable	252
5.5.1 Multiple Regression Analysis: a Brief Introduction	255
5.5.2 Multiple Regression Analysis: Specifications	255

5.5.3	Method of Data Entry	256
5.5.4	Findings (FullA)	257
5.5.5	Findings (IELTSA).....	263
5.6	Interaction (Bias) Analyses.....	265
5.6.1	Interaction Analysis in MFRM: Introduction.....	265
5.6.2	Interaction Analysis: Background Knowledge x Language Proficiency	266
5.6.3	Differential Rater Functioning	269
5.6.3.1	Interaction Analysis: Rater x Topic.....	269
5.6.3.2	Interaction Analysis: Rater x Background Knowledge of Topic	271
5.6.3.3	Interaction Analysis: Rater x General Language Proficiency	271
5.7	Test-Taker Perspectives.....	273
5.8	Summary of Findings	277
Chapter 6	Findings (Qualitative).....	281
6.1	Qualitative Data and Analyses.....	282
6.1.1	Data Sources and Sampling.....	282
6.1.2	Qualitative Analyses.....	284
6.2	Findings	285
6.2.1	Dealing with Topic Unfamiliarity: Topic Abandonment and Test-Taker Strategies.....	285
6.2.2	The Role of General Language Proficiency	287
6.2.3	Topic Unfamiliarity and Negative Affective Influence	288
6.2.4	Topic Unfamiliarity and Impact on Performance and Ratings.....	289
6.2.5	Topic Unfamiliarity and Impact on Rater Decision-Making	290
6.2.6	The Mediating Role of Task Format	291
6.2.7	Problematic Topics and Questions: the Issue of Local Validity	293
6.2.8	Topic Interactions with Characteristics of the Individual Test Taker	298
6.2.9	Test Constraints and Lack of a Support Mechanism.....	300
6.3	Summary.....	302
Chapter 7	Discussion	303
7.1	Summary and Synthesis of Findings	305
7.1.1	Topic Effects on Performance.....	305
7.1.2	Background Knowledge Effects on Performance	307
7.1.3	Background Knowledge and General Language Proficiency	315
7.1.4	Background Knowledge and Topic Difficulty	318
7.1.5	Role of the Topic Development Criterion.....	321
7.1.6	Psychometric Quality of the Speaking Test	322
7.2	Towards Building a Topic Validity Argument.....	323
7.3	Implications	328
7.3.1	Methodological Implications.....	328
7.3.2	Theoretical Implications.....	329
7.3.3	Practical Implications and Recommendations	333
7.4	Limitations.....	336
7.5	Directions for Future Research	338
7.6	Concluding Remarks	340
References		342
Appendices		357

List of Tables

Table 3-1 Illustration of the Incomplete – Connected Data Collection Design in the study-----	98
Table 3-2 Topic-related Questions -----	106
Table 3-3 Rater Profiles-----	119
Table 3-4 The ‘Common Batch’ Rating Design of the Study-----	123
Table 4-1 Summary Statistics for the Analyses Based on the Two Scoring Methods-----	133
Table 4-2 Summary Statistics for the Main C-test Data Analysis-----	144
Table 4-3 Super-item Fit Statistics -----	144
Table 4-4 Overall Item and Person Fit Statistics for the Questionnaire Data-----	151
Table 4-5 Individual Item Fit Statistics for the Questionnaires Items -----	153
Table 4-6 Questionnaire Items Exhibiting Differential Item Functioning -----	157
Table 4-7 Individual Item Fit Statistics for the Questionnaire Items (After Removal of Item 5)-----	158
Table 4-8 Background Knowledge Means of Persons for Each Topic -----	161
Table 5-1 Examinee Measurement Report (Illustrative Examples) -----	173
Table 5-2 Examinees Displaying Overfit or Misfit to the Rasch Model -----	178
Table 5-3 Mean and S.D of Observed and Fair-M Averages for the Different Scales -----	180
Table 5-4 Rater Measurement Report (FullA) -----	182
Table 5-5 Rater Measurement Report (IELTSA)-----	183
Table 5-6 Comparison of Rater Effect Criteria Against the Study’s Rater Fit Statistics-----	187
Table 5-7 Results of the Rack Analyses -----	191
Table 5-8 Rater Severity Estimates in the RackSum Analysis (left) vs. the 3-facet Analysis (right) -	192
Table 5-9 Rating Scale Criteria Measurement Report -----	196
Table 5-10 IELTS Criteria Measurement Report -----	198
Table 5-11 Examinee Range Compared to Topic Difficulty Range Across Analyses -----	200
Table 5-12 Topic Measurement Report (FullA) -----	203
Table 5-13 Topic Measurement Report (IELTSA)-----	205
Table 5-14 Category Statistics: Fluency and Coherence -----	208
Table 5-15 Category Statistics: Lexical Resource-----	211
Table 5-16 Category Statistics: Grammatical Accuracy-----	211
Table 5-17 Category Statistics: Pronunciation-----	211

Table 5-18 Category Statistics: Topic Development-----	211
Table 5-19 Easy vs. Difficult Parallel Forms of the IELTS Speaking Test (FullA) -----	215
Table 5-20 Increase in Average Ability Measures in Adjacent Categories for All Criteria -----	216
Table 5-21 Easy vs. Difficult Parallel Forms of the IELTS Speaking Test (IELTSA)-----	217
Table 5-22 Increase in Average Ability Measures in Adjacent Categories for the IELTS Criteria ---	218
Table 5-23 Task Type A: Topic Measurement Report (FullA)-----	220
Table 5-24 Task Type A: Topic Measurement Report (IELTSA)-----	220
Table 5-25 Task Type B: Topic Measurement Report (FullA) -----	222
Table 5-26 Task Type B: Topic Measurement Report (IELTSA)-----	222
Table 5-27 Task Type C: Topic Measurement Report (FullA) -----	224
Table 5-28 Task Type C: Topic Measurement Report (IELTSA)-----	224
Table 5-29 Increase in Average Ability Measures in Adjacent Categories for All Criteria -----	226
Table 5-30 Increase in Average Ability Measures in Adjacent Categories for the IELTS Criteria ---	228
Table 5-31 Percentage of Observed Functions by Task Type (Illustrative Examples)-----	232
Table 5-32 Average Percentage of Observed Functions Across All Tasks -----	235
Table 5-33 The Background Knowledge Measurement Report (FullA) -----	242
Table 5-34 The Background Knowledge Measurement Report (IELTSA) -----	243
Table 5-35 Increase in Average Ability Measures in Adjacent Categories for All Criteria -----	245
Table 5-36 Increase in Average Ability Measures in Adjacent Categories for the IELTS Criteria ---	245
Table 5-37 MFRM Measurement Result Comparisons (FullA) -----	248
Table 5-38 MFRM Measurement Result Comparisons (IELTSA)-----	249
Table 5-39 Example of a Racked Data Set -----	254
Table 5-40 Regression Model Summary (FullA)-----	258
Table 5-41 ANOVA Output -----	260
Table 5-42 Table of Coefficients -----	263
Table 5-43 Regression Model Summary (IELTSA) -----	264
Table 5-44 Summary of Bias Analysis Results (BK x Proficiency), FullA-----	267
Table 5-45 Summary of Bias Analysis Results (BK x Proficiency), IELTSA -----	267
Table 5-46 Summary of Bias Analysis Results (Rater x Topic), FullA -----	270
Table 5-47 Summary of Bias Analysis Results (Rater x Topic), IELTSA -----	270
Table 5-48 Summary of Bias Analysis Results (Rater x Proficiency), FullA -----	272
Table 5-49 Summary of Bias Analysis Results (Rater x Proficiency), IELTSA -----	272

List of Figures

Figure 1-1 ‘Proficiency’ and its Relation to Performance (McNamara, 1996:86) -----	2
Figure 2-1 Item Characteristic Curves for a Dichotomously Scored Item (i)-----	74
Figure 2-2 ‘A Socio-Cognitive Framework for Test Development and Validation’ (O’Sullivan and Weir, 2011:21)-----	83
Figure 3-1 Task Type A Topics Selected for the Main Study -----	102
Figure 3-2 Task Type B Topics Selected for the Main Study -----	103
Figure 3-3 Task Type C Topics Selected for the Main Study -----	104
Figure 3-4 Data Collection Process -----	128
Figure 3-5 Data Analysis Procedures -----	129
Figure 4-1 Threshold Probability Curves (right) and Category Probability Curves (left) for Items 10, 47 and 264 -----	137
Figure 4-2 Threshold Map for the Texts (Super-items) in the C-tests -----	140
Figure 4-3 Item Characteristic Curves (ICCs) for the Five Easiest Texts (Super-items) -----	141
Figure 4-4 Item Characteristic Curves (ICCs) for the Five Most Difficult Texts (Super-items)-----	141
Figure 4-5 Person-Item distribution Map for the Super-item Analysis-----	143
Figure 4-6 Person-Item Threshold Distribution Map (C-test Main Analysis)-----	145
Figure 4-7 Scatterplot: Speaking Ability Measures vs. C-test Proficiency Measures -----	146
Figure 4-8 Threshold Map for the 8-item Questionnaire Data-----	151
Figure 4-9 Threshold Probability Curves (left) and Category Probability Curves (right) for Item 1--	154
Figure 4-10 Threshold Probability Curves (left) and Category Probability Curves (right) for Item 5	154
Figure 4-11 CPCs for Items 1 (left) and 5 (right) after Collapsing of Categories-----	155
Figure 4-12 Threshold Map of for the Eight Items after Collapsing of Categories)-----	156
Figure 4-13 Item Characteristic Curve for Item 5 (Divided by Proficiency Level)-----	157
Figure 4-14 Item Characteristic Curves for Questionnaire Items 1-4 and 6-8. -----	159
Figure 5-1 A Conceptual-Psychometric Framework of Factors Relevant to this Study (Adapted from Eckes, 2009:11) -----	165
Figure 5-2 Facet Map (4-Facet MFRM) -----	169
Figure 5-3 Person-Item Threshold Distribution (Stack Analysis), left -----	191
Figure 5-4 Person-Item Threshold Distribution (Rack Analysis), right -----	191
Figure 5-5 Category Probability Curves for Raters’ Summed Criteria -----	194

Figure 5-6 Category Probability Curves (Fluency and Coherence)-----	210
Figure 5-7 Category Probability Curves (LR, GA, P, TD)-----	212
Figure 5-8 Facet Map (5-Facet, FullA) -----	240
Figure 5-9 Facet Map (5-Facet, IELTSA)-----	241
Figure 5-10 Distribution of Questionnaire Responses -----	276

List of Appendices

Appendix A. IELTS 9-Band Scale	358
Appendix B. IELTS Speaking Scale (IELTS, n.d.).....	359
Appendix C. Task Equivalence Checklist	360
Appendix D. Task Statistics	361
Appendix E. Speaking Test Form W.....	364
Appendix F. Speaking Test Form X.....	366
Appendix G. Speaking Test Form Y.....	368
Appendix H. Speaking Test Form Z	370
Appendix I. C-tests	372
Appendix J. Invitation (Language Schools).....	378
Appendix K. Participant Information Sheet	379
Appendix L. Participant Consent Form.....	381
Appendix M. Invitation (Raters)	382
Appendix N. Topic Development Scale (ETS, 2004).....	383
Appendix O. Observation Checklist (Additional Details)	384
Appendix P. Sample Score Sheet.....	385
Appendix Q. Score Sheets: Instruction to Raters	386
Appendix R. Examinee Measurement Report.....	388
Appendix S. Table of Unexpected Responses	390
Appendix T. Multiple Regression (Assumptions)	393

Abbreviations and Acronyms

ACTFL	American Council on Teaching of Foreign Languages
ANOVA	Analysis of Variance
BK	Background Knowledge
CA	Conversation Analysis
CEFR	Common European Framework of Reference
CPC	Category Probability Curve
CRHE	Common Rater Halo Effect
DIF	Differential Item Functioning
EAP	English for Academic Purposes
ESL	English as a Second Language
EFL	English as a Foreign Language
ESOL	English for Speakers of Other Languages
FC	Fluency and Coherence
FullA	Analysis with All criteria
GA	Grammatical Range and Accuracy
ICC	Item Characteristic Curve
IELTS	International English Language Testing System
IELTSA	Analysis with IELTS Criteria
IRT	Item Response Theory
IST	IELTS Speaking Test
ITA	International Teaching Assistant
L1	First Language
L2	Second Language
LR	Lexical Resource
LSAP	Language for Specific Academic Purposes
LT	Language Testing
MFRM	Multi-Faceted Rasch Measurement
P	Pronunciation
PSI	Person Separation Index
RMSE	Root Mean-Square Error
SD	Standard Deviation
SE	Standard Error
SLA	Second Language Acquisition
T-LAP	Test of Listening for Academic Purposes
TOEFL	Test of English as a Foreign Language
TOEFL iBT	Internet-based Test of English as a Foreign Language
TPC	Threshold Probability Curve

Chapter 1 Introduction

In performance-based assessments of spoken language, a common practice for eliciting samples of speech is to engage test takers with speaking tasks on a given topic or range of topics. In order to address these topics, test takers often draw on their topic-related background knowledge which generally serves as an information base for performance to be built upon (Bachman and Palmer, 1996). In administering different topics to test takers, there is an underlying assumption that regardless of the choice of topic, the speaking tasks are of equivalent levels of difficulty and can be considered ‘parallel’. What logically follows is a second underlying assumption that the individual test takers’ degree of topic-related background knowledge does not have a significant influence on test results.

However, evidence to the contrary may suggest that a validity threat has been introduced to the test owing to the influence of the construct-irrelevant factor of background knowledge of topic. Moreover, test fairness may also be compromised if it is shown that individuals or groups of individuals have been favoured or biased against as a function of their background knowledge of topics (Jennings et al. 1999). Given that the results of tests, particularly large scale standardised ones, are used to make decisions about test takers, these validity concerns become critical. However, a review of the literature on the effects of topic and background knowledge of topic on performance points to a surprising dearth of empirical research which has touched on these issues. This thesis, in response, aims to examine the effects of the two variables of topic (a task characteristic) and background knowledge of topic (a test-taker characteristic) on spoken performance scores. The next section provides a brief overview of the parameters involved in testing spoken

performance in order to position the two variables of interest within the larger and more complex context of oral test performance.

1.1 Variability and Spoken Test Performance

Variability in second language spoken performance assessment is an area of significant interest and debate from both a theoretical and an empirical standpoint. It has been posited that myriad factors other than the speaking ability in question may have the potential to influence performance in an assessment context. McNamara's (1996:86) model of 'proficiency' reproduced in Figure 1-1 captures some of the complexities involved in test performance¹.

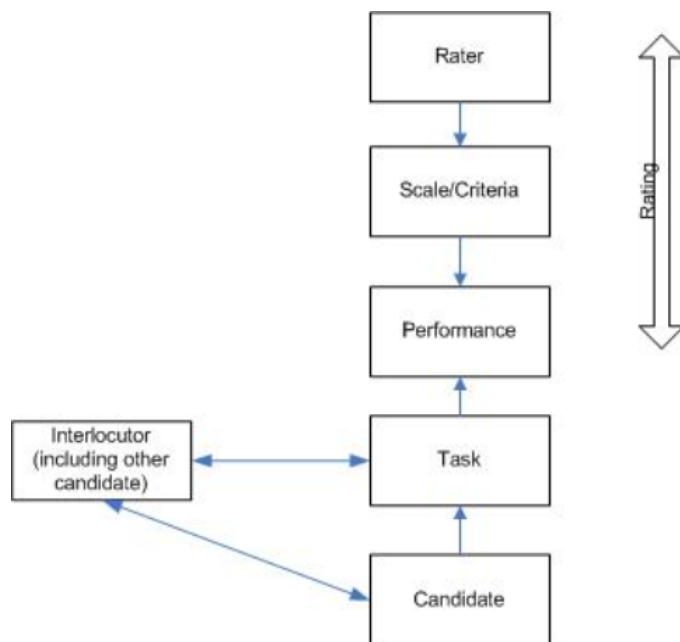


Figure 1-1 'Proficiency' and its Relation to Performance (McNamara, 1996:86)

¹ This is a model of 'direct' performance assessment which involves the candidate in direct interactions with human interlocutors and raters, and should be distinguished from semi-direct or fully automated spoken performances which involve the use of computers for test delivery and/or scoring.

As the figure illustrates, spoken performance can be influenced by several parameters: characteristics of the candidate, the task, the interlocutor, the rater, the rating scale and criteria, as well as the interactions between them.

Kunnan (1995:6), for example, lists several test-taker characteristics such as age and gender as well as other social, cognitive and psychological factors such as cultural background, aptitude and learning styles as having a potentially ‘critical influence’ on second language performance (see also O’Sullivan, 2000a). Other factors which have been identified in the literature as potentially exerting an impact on performance include, but are not limited to, the characteristics of the interlocutor(s) and/or rater(s) for example, gender, proficiency level, personality (O’Sullivan, 2000b; Nakatsuhara, 2011), raters’ degree of harshness or leniency when rating (McNamara, 1996; Eckes, 2009), raters’ approaches to scoring and interpretations of rating scale criteria (Milanovic, Saville and Shuhong, 1996; Pollitt and Murray, 1996; Lumley, 2002; Cumming, Kantor and Powers, 2002, Baker, 2012), characteristics of the rating scales (Barkaoui, 2007; 2010), as well as socially constructed phenomena such as power relations (Shohamy, 2001) and degree of acquaintanceship (O’Sullivan, 2002). Speaking test tasks, as the instruments used to elicit spoken performance, have also been extensively studied in the literature. It has been suggested that task features and processing conditions can be manipulated to guide test takers’ performance on the one hand and to vary task difficulty on the other (Luoma, 2004; Skehan and Foster, 1997).

The reason why variability in spoken assessment has attracted so much attention in the literature is due to the fact that the resulting score from a test is used for making inferences about a test taker’s abilities and also for making important decisions about them.

However, as McNamara (2000:55) points out, a score can be ‘attractively simple’ and yet ‘deceptive’. To exemplify, if a particularly harsh rater assigns a score to a candidate, it may not be clear whether the score reflects the candidate’s ability or whether it is confounded with the rater’s severity. Similarly, if a test taker is asked to talk about a topic about which they have little background knowledge, it will be difficult to interpret the score, as it might not be clear whether it reflects speaking ability or whether the score has been negatively influenced by the test taker’s lack of background knowledge. From a test validity perspective, it is therefore important to monitor the potential effects of factors which are extraneous to the ability being measured, that is, construct irrelevant variables (McNamara, 2000), and to consider ‘other plausible rival interpretations of score meaning’ (Messick, 1996:246). We now return our attention to the factors of interest in the current study.

1.2 Speaking Task Characteristics: Topic

Speaking test tasks play a pivotal role in assessment; they serve as a link between test takers’ underlying abilities and subsequent performance by eliciting samples of speech from test takers (Fulcher, 2003). Speaking tasks can be defined as ‘activities that involve speakers in using language for the purpose of achieving a particular goal or objective’ within particular settings (Bachman and Palmer, 1996:44). Moreover, by manipulating task characteristics and administration conditions, test designers can direct and influence candidates’ performance to a certain extent (Luoma, 2004). Relevant to this discussion is a distinction made by Brown et al. (1984) between ‘chatting’ and ‘information-related talk’ as representing two ends of a continuum in respect to the purposes of ‘talk’. Chatting is viewed as a predominantly social activity which involves ‘finding a fluid stream of

topics that the speakers find sufficiently interesting to take up, and on which they can find a shared angle. The topics are not necessarily discussed very deeply' (Luoma, 2004:22). At the other end of the continuum is 'information-related talk' which is described as 'speech aimed at transferring information on a particular topic' (Luoma, 2004:23) and is the one more often used in assessment contexts (*Ibid.*). The information-oriented nature of speaking tasks in assessment contexts thus highlights the importance of the task topic and the test takers' information about the specific topic.

As a task characteristic that can influence performance, topic features heavily in models of language use and task-based performance. In Bachman and Palmer's (1996) influential model of communicative language ability, topic is identified as a component of the language of test task 'input' with input described as what the test takers are supposed to process and subsequently respond to. Other components of the input include 'format (e.g. channel, form, length, type) and language (organizational, pragmatic and topical characteristics)' (Bachman and Palmer, 1996: 52-53). The topic component carries information in the input of the task and can be 'personal, cultural, academic, or technical' (Bachman & Palmer, 1996:53). The potentially facilitating or impeding role of topical knowledge in relation to task topics is commented on by the authors, who argue that 'certain test tasks that presuppose cultural or topical knowledge on the part of test takers may be easier for those who have that knowledge and more difficult for those who do not' (Bachman and Palmer, 1996:65). Similarly, in Skehan's (1998) framework of task processing conditions, a facilitative role for topic familiarity on performance has been suggested. It is hypothesised that the more familiar a topic, the less cognitive load it poses on the individual by providing 'easy access to information [which] should make only

limited demands on attention, allowing material to be assembled for speech more easily and with greater attention to detail' (Skehan, 2001:175). Weir (2005:76) also voices a concern that different topics may elicit 'responses that are *measurably* different'.

While topic is viewed as an important characteristic of tasks in general, it is logical to assume that its salience may also be affected by task type. A discussion of various classification systems for tasks is beyond the focus of this thesis (see Weir, 1988 for more information). However, attention is drawn to two specific task types which are germane to this discussion: *integrated speaking tasks* and *independent* or *stand-alone* speaking tasks. Integrated speaking tasks are defined as tasks that 'involve combinations of reading, listening and/or writing activities with speaking' (Luoma, 2004:43) and require test takers 'to speak about a topic for which information [has] been supplied from another source' (Jamieson et al., 2008:76). In contrast, independent or stand-alone tasks require test takers to 'draw on their own ideas or knowledge [in order] to respond to a question or prompt' (Brown, Iwashita and McNamara, 2005:1).

These two task types have been compared on several aspects such as degree of authenticity (particularly in academic contexts), content coverage, generalisability, cognitive processing demands on test takers and reliability of ratings (Lewkowicz, 1997; Luoma, 2004; Huang, 2010). However, it is to their different approaches to addressing topic-related background knowledge that attention is drawn. In integrated tasks, an attempt is made to minimise and/or mediate the (negative) impact of background knowledge through the provision of input in the form of reading and listening texts (Read, 1990). In contrast, independent tasks require test takers to draw on their own background knowledge. This absence of input in independent tasks has been criticised for not allowing

an ‘equal footing in terms of background knowledge’ (Weigle, 2004: 30) for test takers who bring varying degrees of background knowledge to a test (Huang, 2010), and for the restriction of topics to ‘fairly bland’ ones (Brown, Iwashita and McNamara, 2005:1). In this light, integrated tasks are viewed as ‘promoting equity or fairness’ (Huang, 2010: 4). It can also be argued that by providing test takers with the necessary ideas for responding to a topic, as opposed to asking them to generate ideas and to fall back on their background knowledge, the cognitive demand of tasks can be reduced to a certain extent (Field, 2011). As Jennings et al. (1999) caution, this is not to say that a topic effect does not exist in integrated tasks but that the impact is likely to decrease owing to the provision of input. By the same token, it is plausible to assume that any effects of topic and background knowledge of topic are likely to be manifested more markedly in independent tasks.

1.3 Test-taker Characteristics: Background Knowledge of Topic

As evident in the previous section, a discussion of task topics is inextricably linked to test takers’ background knowledge of topics. To clarify, background knowledge (BK) is also referred to in the literature under different terms such as topical knowledge (Bachman and Palmer, 1996), prior knowledge (Tedick, 1990), world knowledge (O’Sullivan and Green, 2011), and topic familiarity (Skehan, 1998) and has been identified in the theoretical literature as having a substantial effect on performance. It features as one of the five main components of Bachman and Palmer’s (1996) model and is viewed as inseparably linked to all instances of language use, as ‘it provides the information base that enables them [individuals] to use language with reference to the world in which they live, and hence is involved in all language use’ (Bachman and Palmer, 1996:65). Background knowledge is

often considered in relation to potential sources of test bias if the test task ‘contains content or language that is differentially familiar to subgroups of test takers’ (O’Sullivan and Green, 2011:61).

1.4 Statement of the Problem

In spite of the pronounced role attributed to topic and background knowledge of topic on performance in a second language from a theoretical standpoint, the results of empirical studies are often mixed and inconclusive. One possible reason is the various ways in which background knowledge has been operationalised in the literature, for example, as cultural background (He and Shi, 2012), academic field of study (Clapham, 1996), religious background (Markham and Latham, 1987), personal interest in topics (Jennings et al., 1999) and gender (Lumley and O’Sullivan, 2005). Furthermore, the majority of empirical studies have explored the effects of topic and BK on reading and listening skills while minimal attention has been paid to the performance skills of writing and speaking, with only a handful of studies exclusively focusing on topic effects on speaking. This is surprising, as the case for speaking is arguably stronger than the other skills; the online nature of speaking requires almost instantaneous access to background knowledge for the spontaneous generation of ideas necessary for addressing topic-based tasks. In light of the paucity of research on this subject and the importance attributed to the two associated factors of topic and background knowledge of topic, a close examination of these two variables on speaking performance is needed and constitutes a rationale for the current study.

A second rationale for the study is directly linked to the specific speaking test selected as the context of this thesis: the Speaking Module of the International English Language Testing System (IELTS). It is argued that despite the centrality of topic as a speech-elicitation method in this test, the topic effect has rarely been addressed in relation to the speaking test whereas one would expect an examination of test topics to be a priority from a test validation perspective, given the high stakes nature of the test.

The next section provides some general background to the IELTS test before providing a detailed description of the Speaking Test. The section also serves as a frame of reference for comparing the differences and similarities between the experimental setting of the current study and the operational setting of the IELTS Speaking Module. This will be particularly pertinent when generalising findings and drawing conclusions about the operational test. Once the Speaking Test has been described, the rationale for selecting the test as the context of the study is re-visited.

1.5 What is IELTS?

The International English Language Testing System (IELTS) is a high stakes standardised test of English language proficiency. It has a candidature of approximately 1.7 million per year and is taken in 800 test centres in more than 130 countries around the world (IELTS, 2012). IELTS is jointly managed by three organisations: the British Council, IDP: IELTS Australia² and the University of Cambridge ESOL Examinations.

² IDP is an independent international Australian organisation for educational and professional development.

The Academic Module and the General Training Module are the two test versions available to candidates³. The former is designed to measure English language proficiency required for academic settings – although not subject-specific – and is more extensively used as an admissions criterion for higher education institutions whereas the latter is designed to measure proficiency for more practical use in social contexts and at the workplace, and is used by governmental and other organisations for immigration purposes, training, secondary education or work in English-speaking countries (IELTS, 2011). Both versions assess the four language skills (Listening, Reading, Writing and Speaking) and the only distinguishing factor between the two is the subject matter of the Reading and Writing modules. The Listening and Speaking Modules are the same for both versions. IELTS scores are reported on a 9-band scale, designed to distinguish language ability levels from ‘non-user to expert user’ (see Appendix A). There is no pass or fail criterion in IELTS on the grounds that ‘the level of English needed for a non-native speaker student to perform effectively varies by situation and institution’ (IELTS, 2011:13). More details about the test can be found on the IELTS website (<http://www.ielts.org>). For information about the historical development of IELTS, the interested reader is referred to Davies’ (2008) excellent book, which chronicles the evolution of IELTS in the format it is known today against the historical backdrop of academic English language proficiency testing within the UK higher educational system. In reflecting on the IELTS test as it currently stands, Davies (2008) draws attention to the ongoing conflict between theoretical stances in defining constructs of language proficiency within applied linguistics on the one hand, and the practical demands of standardised large-scale assessment on the other. He

³ In 2011, the split between the candidature across the two versions was 78.3% and 21.7% for Academic and General Training respectively (IELTS, 2012).

describes this balancing act as ‘reconciling theoretical developments in applied linguistics and language testing and in measurement theory with a testing organisation’s requirements of production and delivery’ (Davies, 2008: 93). This is an important point to bear in mind, as practicality concerns, particularly those associated with large-scale assessment, are rarely touched on in experimental studies.

As noted earlier, IELTS scores are used by different organisations to make important decisions about individuals’ future academic and professional lives, which is why IELTS is categorised as a high-stakes test: ‘high-stakes decisions are major, life-affecting ones where decision errors are difficult to correct. Because of the importance of their effects, the costs associated with making the wrong decision are very high’ (Bachman, 2004:12). Similarly, Shohamy (2001:102) draws attention to test takers and defines the criterion for a high-stakes test as ‘whether the results of the test lead to detrimental effects for the test takers’. It is this very high-stakes nature of IELTS which demands rigorous research on every aspect of the test in order to ensure its validity and fairness. Cronbach (1988:6) argues that ‘tests that impinge on the rights and life chances of individuals are inherently disputable’ and they should thus be accessible to critical reflection and dialogue within the language testing community (Fulcher and Davidson, 2007). The IELTS partners’ response to these fairness and validity concerns has been to establish ongoing internal and external research programmes to address such issues of test validity. They have also made available online some – albeit limited – access to materials, data and resources which makes research on the test a viable option for non-IELTS funded projects, including the current study.

1.6 The IELTS Speaking Test

1.6.1 Test Format, Duration and Settings

The IELTS Speaking Test (IST) is a face-to-face oral interview between a test taker and a certified IELTS examiner. The interview is comprised of three parts, lasts between 11-14 minutes and is recorded⁴. The test taker usually arrives at a prearranged location such as a test centre at a specific scheduled time and is directed to a quiet room where the speaking test is administered by the examiner.

1.6.2 Test Parts, Topic Sets and the Examiner Script

There are three main parts in the IELTS Speaking Test. Part 1 (4-5 minutes) consists of an *Introduction* where the examiner introduces himself/herself, asks the candidate to introduce himself/herself and confirms the candidate's identity, and is followed by an *Interview* task (also known as *Information Exchange* task) where the examiner poses a series of questions on some general and familiar topics. In Part 2 of the test, the *Individual Long Turn*, the candidate is presented with a printed task card, which requires the candidate to give an extended talk on a specific topic for 1-2 minutes. The task card includes points that the candidates can cover in their monologue. Prior to the monologue, the candidate is given a one-minute silent planning time as well as pencil and paper for taking notes. Once the long turn is finished, the examiner may round off the topic with one or two questions. This part takes about 3-4 minutes. The third part of the test, which lasts about 4-5 minutes, is termed a *Two-way Discussion* where the examiner poses several questions on more abstract topics, which are thematically linked to the topic in Part 2. This

⁴ The speaking tests are recorded for monitoring purposes and re-marking (in case of a results enquiry).

final part aims to provide the candidate with an opportunity to discuss these more abstract themes and topics.

Part 1 (*Interview* or *Information Exchange*) consists of three multi-question topic sets (also known as topic *frames*). The first set involves a general topic such as work or studies and the remaining two sets involve other familiar topics. Part 2 (the *Individual Long Turn*) consists of one topic-based monologue task. Similar to the *Interview* task, the *Two-Way Discussion* task often consists of two topic sets. It should be noted that the topic sets for Parts 1 and 2 are randomly assigned to test takers and the topic sets from Part 3 of the test are linked to the assigned topic in Part 2. The *Interview (Information Exchange)*, *Individual Long Turn*, and *Two-way Discussion* task types corresponding to the three parts of the test are henceforth referred to as Task Types A, B and C respectively.

Prior to the live tests, the speaking task materials undergo rigorous quality control checks. Taylor (2007a: 25-28) describes the production cycle of the IELTS Speaking and Writing materials known as the 'IELTS Question Paper Production or QPP' and explains how both qualitative and quantitative procedures are set in place in order to ascertain the 'careful selection of appropriate test material and the maintenance of consistent levels of test difficulty over time'. The QPP is comprised of several stages of 'task commissioning, pre-editing, editing, (...) trialling (with a representative sample of the test taking population) and review' (Taylor, 2007a: 26). Once finalised, the speaking tasks 'are located in the test construction bank to await compilation into live sets of material' (*Ibid.*:26). The pre-editing stage specifically targets topic and topicality issues to ensure that 'all test material is suitable for editing, is culturally appropriate and accessible worldwide' (Taylor, 2007a: 27).

The IST is scripted and standardised, and examiners are given detailed instructions to follow in terms of test administration and management. The reliability of test delivery is achieved through the use of an *Examiner Frame* or *Script* which is ‘**a script that must be followed** (original emphasis)’ (IELTS Examiner Training Material 2001: 5 cited in Seedhouse and Harris, 2011: 72)⁵ and which is designed to carefully delineate the examiner’s role in the interaction with the candidate, ‘[guiding] test management’ as it progresses through different parts of the test (Taylor, 2007b: 187). The examiner script was developed in response to research findings which provided evidence of a significant effect of examiner variation – in terms of interactional style – on candidate performance in the IELTS Speaking Test (Brown and Hill, 1998).

1.6.3 Test Construct, the Rating Scale and the Rating Process

The construct underlying the Speaking Test is communicative (spoken) language ability. The test is designed to assess a wide range of skills which correspond to the different parts of the test and aims to elicit ‘the ability to communicate opinions and information on everyday topics and common experiences and situations by answering a range of questions; the ability to speak at length on a given topic using appropriate language and organising ideas coherently; and the ability to express and justify opinions and to analyse, discuss and speculate about issues’ (IELTS, 2010: 4). Each part of the test ‘fulfils a specific function in terms of interaction pattern, task input and candidate output’ (Seedhouse and Harris, 2011: 72). The speech functions typically observed in the test include providing personal and non-personal information; comparing; contrasting; explaining; expressing and justifying opinions; narrating; expressing preferences;

⁵ The IELTS training material are strictly confidential. This is why secondary sources are used when referencing these materials.

summarising; paraphrasing; conversation repair; suggesting; analysing and speculating (Brooks, 2003).

The IELTS speaking scale is a 9-band analytic scale consisting of four criteria: *Fluency and Coherence* (FC), *Lexical Resource* (LR), *Grammatical Range and Accuracy* (GA) and *Pronunciation* (P). For each criterion, the expected spoken performance at each level is defined in the band descriptors (see Appendix B for the public version of the speaking band descriptors). A succinct summary of the four criteria and their key indicators were outlined in Seedhouse and Harris (2011: 72-73) and are reproduced below:

‘Fluency and Coherence refers to the ability to talk with normal levels of continuity, rate and effort and to link ideas and language together to form coherent, connected speech. The key indicators of fluency are speech rate and speech continuity. For coherence, the key indicators are logical sequencing of sentences, clear marking of stages in a discussion, narration or argument, and the use of cohesive devices (...)within and between sentences.

Lexical Resource refers to the range of vocabulary the candidate can use and the precision with which meanings and attitudes can be expressed. The key indicators are the variety of words used, the adequacy and appropriacy of the words used and the ability to circumlocute (...) with or without noticeable hesitation.

Grammatical Range and Accuracy refers to the range and the accurate and appropriate use of the candidate’s grammatical resource. The key indicators of grammatical range are the length and complexity of the spoken sentences, the appropriate use of subordinate clauses, and variety of sentence structures, and the ability to move elements around for information focus. The key indicators of grammatical accuracy are the number of grammatical errors in a given amount of speech and the communicative effect of error.

Pronunciation refers to the capacity to produce comprehensible speech in fulfilling the Speaking Test requirements. The key indicators will be the amount of strain caused to the listener, the amount of unintelligible speech and the noticeability of L1 influence.’

Scores are awarded for each criterion (as whole bands) and the four scores across criteria are subsequently averaged and rounded to the nearest upper half band or whole band score (IELTS, 2012). It is this average which is reported as the speaking band score on IELTS test report forms.

IELTS examiners evaluate candidate performances throughout the test and award a final score once the candidates have left the session. The IELTS examiner thus serves the dual role of interlocutor and rater. The scoring of the Speaking Test follows a single-marking model, that is, only one score is awarded per performance and double marking is not carried out for the test⁶. Test reliability is assured through examiner training and certification (as well as re-certification every two years), standardised procedures, a sample monitoring system⁷, a ‘jagged profile’ system⁸ and routine validation of task, candidate and examiner performance (Taylor, 2007a: 29).

1.6.4 The IELTS Speaking Test: a Critical Review

So far, the discussion on the IELTS Speaking Test has been predominantly descriptive. In this section, the test is reviewed from a critical perspective with an eye for contextualising the previous discussions on topic effects on performance within the specific features of the Speaking Test. The review draws extensively on a recent work by Seedhouse and Harris (2011), which focuses on topic development in the IST. Seedhouse and Harris (2011) applied Conversation Analysis (CA) institutional discourse methodology⁹ to a corpus of 60 audio-recorded spoken performances which covered a range of proficiency levels from

⁶ Unless a formal enquiry has been made about a mark.

⁷ A sample monitoring process ‘requires selected centres worldwide to provide a representative sample of examiners’ marked tapes and scripts such that all examiners working at a centre over a given period are represented (...) Tapes and scripts are then second-marked by a team of experienced IELTS examiners worldwide. Examiners monitor for quality of both test conduct and rating, and feedback is returned to each centre for follow-up with individual examiners’ (Taylor, 2007a:29)

⁸ A ‘jagged profile system’ involves ‘routine targeted double marking of candidates identified as being at risk of misclassification, based on the level of divergence between their Writing and/or Speaking scores and their Reading and Listening scores’ (Taylor, 2007a:29)

⁹ The aim of CA institutional discourse is to describe the interaction in a test but also to relate this interaction to institutional goals: ‘CA institutional discourse methodology attempts to relate, not only the overall organisation of the interaction, but also individual interactional devices to the core institutional goal. CA attempts, then, to understand the organisation of the interaction as being rationally derived from the core institutional goal’ (Seedhouse and Harris, 2011: 74).

IELTS band 4.0 to 9.0 in order to examine the management and development of topic in the Speaking Test. The findings from this study are referred to throughout the discussion.

Let us now revisit the structure of the IST. In all three parts of the test, speech is elicited by means of different topics, from more familiar ones in Part 1 to more abstract themes in Part 3. This illustrates the heavily topic-based nature of the test. The findings from the conversation analysis of transcribed IELTS spoken performances in Seedhouse and Harris (2011) also identified topic as an integral component of the IST and as a vehicle for organising talk, driven almost exclusively by the examiner script. Referring to the organisation of talk in the test as a ‘topic-based Q-A adjacency pair’, the authors illustrate, in their examination of representative performances, how, in contrast to normal conversation, ‘topic is always introduced by means of a question’ (Seedhouse and Harris, 2011: 69). The two elements of the adjacency pair involve a question posed by the examiner to which the candidate has to respond and a ‘topic’ element which calls for the development of a specific topic (Seedhouse and Harris, 2011: 83). The findings also suggested the primacy of the Q-A element over the topic element in those instances where the two do not co-occur as ‘candidates can answer questions without developing topics’ (*Ibid.*:83). Put differently, the provision of a response, which might only minimally answer a question may be perceived by candidates to be more important than elaborating on a response by means of topic development. In these cases, the information-transfer function of the test takes precedence over its main purpose, that is, speech generation for the purposes of evaluation.

A careful perusal of the Speaking Test band descriptors (Appendix B) reveals the absence of a topic development or content-oriented criterion which might explain why test takers

may not always elaborate on topics (Seedhouse and Harris, 2011), as the extent to which ideas are developed has not been explicitly defined as part of the speaking test construct definition (Sato, 2012). Topic development only features at the higher levels of the scale to distinguish bands 8 and 9 under *Fluency and Coherence*. Topic-related vocabulary is also a key indicator under the *Lexical Resource* scale. However, there is no independent topic development scale which could capture the extent to which candidates are successful in developing topics and/ or completing the tasks.

The inadequacy of the current rating scales in dealing with ‘off-topic responses’ or in addressing limited topic development is documented in Brown and Taylor’s (2006:16-17) survey of IELTS examiners, as a result of which the researchers suggested the inclusion of a ‘task-response scale’ or a ‘specific focus on how the candidates answer the questions’. This recommendation has not been applied thus far, possibly owing to the fact that any such scale would have to be task/topic-dependent. However, performances on different tasks and parts in IELTS are not separately scored, but are given a final rating once all tasks have been completed. Such an approach has been criticised by O’Sullivan and Nakatsuhara (2011:182) who argue that ‘a single score for performance on a number of tasks does not offer a true reflection of a candidate’s true ability’ and ‘in many different tests, where a single score is awarded for performance on a set of tasks, we can never be certain what has influenced the examiner’s decision to award a particular score’ (*Ibid.*:183).

The decision by IELTS developers to not award separate scores per task and/or not include a topic development/task-response scale is perhaps due to practical constraints and the increased cognitive demand for examiners of such an approach, where they would not

only have to manage the test but also rate each task as they move through different stages of the test. Nevertheless, the role of topic development in allowing the test to achieve its main purpose of generating samples of ratable speech requires careful consideration regarding whether a content-oriented criterion (which emphasises the development of ideas and topics) should be explicitly included in the scales (Sato, 2012).

Two other aspects of the test will be touched on as factors that can amplify the topic effect. Firstly, the fact that the IELTS speaking tasks are not integrated tasks, but independent, and thus rely on test takers bringing their own background knowledge in addressing the task topics and answering topic-related questions. While it can be argued that by adopting a multi-task structure and by including a mix of topics from more familiar to more abstract (and by providing support in the form of both planning time and ideas in the second part of the test) the test developers have aimed to obviate or minimise the topic effect, test takers still need to draw on their background knowledge and deal with the online spontaneous demands of addressing different topics. Secondly, the level of control exerted by the examiner on the test and the ‘asymmetrical rights to topic management’ (Seedhouse and Harris, 2011:15) are factors which can impede test takers from shifting topics that they find problematic as the ‘management of topic is almost entirely determined by the examiner’s script’ (*Ibid.*:15). While Seedhouse and Harris (2011) identify the dominance of the examiner script as being in line with the institutional goals of achieving reliability and standardisation, the combination of absence of choice for test takers and unequal power dynamics which, taken together, do not allow candidates to shift or navigate topics can intensify the potentially negative effects of topic when test takers do not have the background knowledge necessary to address the topics or tasks.

At the same time, and from a research perspective, the use of the examiner script which aims to minimise the interlocutor effect, makes it possible to isolate the topic effect on speaker performance to a large extent whereas a less-structured test format and/or a paired or group format would have introduced myriad confounding variables in examining topic effects.

In sum, the above discussion illustrates how the effects of topic can be of particular salience in the IELTS Speaking Test on the following grounds: the test consists of non-integrated independent, heavily topic-based tasks with an information-transfer oriented purpose which require test takers to rely on their background knowledge to respond to questions and develop topics. Moreover, the test takers are not allowed any choice in topic selection and have little control in topic management. It therefore makes intuitive sense that the topic of the task and the background knowledge that test takers bring to the task may exert a substantial positive or negative influence on performance. On the other hand, the one-to-one interview format and the scripted and standardised nature of the test, makes the IST a viable option for examining the topic effect in relative isolation from interlocutor-related factors in paired or group interview formats. The various strands of discussion addressed above are now brought together in presenting the rationale for the study.

1.7 Rationale for the Study

Shohamy (2001) highlights the importance of a biographical depiction of a researcher's personal and professional journeys in achieving a greater understanding of their perspectives and rationales for research. My interest in language testing in general, and in

the above field of inquiry in particular, stems from experience as an IELTS candidate, a teacher of IELTS for academic purposes, and, finally, as an IELTS examiner. There were numerous instances of anecdotal evidence which pointed to the candidates' struggling with some of the IELTS speaking test topics, leading them to claim that they 'had nothing to say' about the topics, regardless of their speaking proficiency. It almost makes intuitive sense for topic (and having something to say about a topic) to make a difference on performance even in one's first language let alone a second language and particularly, in a high-stakes test setting. It was a combination of such observations and test-taker concerns that first triggered in me an interest in this important area of validity inquiry.

Secondly, from a theoretical perspective, it has been posited that both topic and background knowledge of topic can have an impact on performance: topic, in providing a temporary meaningful context for the elicitation of ratable samples of speech; and background knowledge of topic, in providing the information base on which speech is built (Bachman and Palmer, 1996).

Thirdly, empirical studies which have examined the effects of the two variables of interest on performance are scarce and their results inconclusive. The foci of the majority of these studies have been predominantly on the receptive skills of listening and reading with very few studies involving the productive skills of writing and speaking. Apart from a select few studies which have looked at topic effects in integrated contexts (e.g. Jennings et al., 1999 on writing; Huang, 2010 on speaking), no empirical study, to the best of my knowledge, has explored the effects of the two variables of interest in the context of a non-integrated, independent high-stakes test of general speaking proficiency; a context which arguably can be most susceptible to a (negative) topic effect. As illustrated earlier, the IST

is a prime example of such a high-stakes test where the potential topic effect constitutes a validity threat to the test (Messick, 1989).

Fourthly, the random assignment of tasks and topics to test takers in the IST implies an underlying assumption that speaking tasks, in spite of having different topics, are of equivalent levels of difficulty. According to Weir (2005), it is critical to empirically demonstrate the equivalence of such test tasks in order to ensure fairness to candidates. Moreover, an examination of the test tasks and topics is encouraged as means of ensuring ‘critical reflection and dialogue’ (Fulcher and Davidson, 2007:61) amongst the test’s stakeholders. Taken together, these reasons constitute the rationale for the proposed study in addressing a substantial gap in spoken performance assessment research which can also have important practical implications for one of the major English language tests in the world.

1.8 Research Questions

The study’s main research question and subsidiary research questions are designed to explore the effects of topic and background knowledge of topic on spoken performance in light of the requirements of test validity, and in reference to Weir’s (2005) socio-cognitive framework of language test validation (see section 2.4.3).

The study’s main overarching research question is formulated as follows:

How is the topic validity of a test of second language speaking performance influenced by the random assignment of topics to test takers who bring different levels of background knowledge to the topics?

In order to address this research question, the study adopted both quantitative and qualitative research methods. The following subsidiary research questions guided the collection of different sources of validity evidence:

- i. To what extent are the topics (speaking tasks) used in parallel versions of a language proficiency interview similar in terms of difficulty?*
- ii. To what extent does the observed progression of topic difficulty measures match the intended progression of topic difficulty from easy to difficult?*
- iii. To what extent are parallel forms of a language proficiency interview (consisting of different topics) comparable in terms of difficulty? Are (any) differences in form difficulty large enough to have practical significance in terms of test performance?*
- iv. When task type is held constant, to what extent are the different topics used in parallel versions of a task similar in terms of difficulty? Are (any) differences in topic difficulty measures large enough to have practical significance in terms of test performance?*
- v. When task type is held constant, to what extent are the observed functions elicited by different topics similar?*
- vi. Will the differences in test takers' levels of background knowledge of topics have an impact on performance? Are (any) differences large enough to have practical significance in terms of test performance?*
- vii. Does background knowledge of topics have an impact on topic difficulty measures?*
- viii. Does background knowledge of topics differentially affect performances of test takers from different proficiency levels?*
- ix. Are there any significant interactions between raters and a) topics, b) test takers' background knowledge of topics and c) test takers' language proficiency levels?*
- x. What are the test takers' perceptions of the influence of topic and background knowledge of topic on performance?*

xi. What themes and patterns emerge from a qualitative inquiry into a) rater perspectives and b) the content of test-taker speaking performances, in relation to topic validity aspects of the speaking test under examination?

1.9 Structure of the Thesis

This introductory chapter identified the variables of interest for the present study, introduced the specific test under examination, presented the rationale behind the study, and specified the study's main and subsidiary research questions.

Chapter 2 contextualises the study within the broader literature by providing a critical review of studies which have investigated the effects of the two variables of topic and background knowledge of topic on performance, as well as studies pertaining to the influence of raters in performance assessment contexts. Next, the chapter presents an introduction to the Rasch model as an appropriate method for addressing the complexities involved in performance assessment. The chapter ends with an overview of the current understandings of validity and validation followed by a description of the language test validation framework adopted in the study.

Chapter 3 details the study's research methodologies in terms of data collection and analysis. It provides a detailed account of the specific instruments used for different purposes e.g. task selection (task equivalence checklist), data collection from participants (e.g. speaking task topics, language proficiency tests, background knowledge questionnaires) and data collection from raters (rating scales, observation checklists). Justifications are made for the methodologies opted for in light of the literature review as well as theoretical and practical considerations. Issues and concerns regarding these

decisions are raised where relevant. The particulars of procedures for quantitative and qualitative data analyses are described.

Chapter 4 presents the results of the preliminary analyses of data from two of the study's research instruments designed to measure general language proficiency and background knowledge of topics respectively. This chapter provides information regarding the quality and functioning of the instruments. The resultant data are used in running the study's principal analyses.

Chapter 5 reports the results of the quantitative analysis of data from Multi-Faceted Rasch Measurement in relation to the study's main facets i.e. examinees, raters, criteria, topics and background knowledge of topics. Results of a multiple regression analysis are drawn upon to specify the predictive power of the two main variables of background knowledge of topics and general language proficiency. Findings of bias analyses between pre-selected facets are also provided. The findings from test takers' close-ended questionnaire responses are reported last.

Chapter 6 presents the results of a qualitative analysis of the content of test takers' spoken performances in order to gain an insight into possible sources of interactions between test takers and topics. The findings from the qualitative analyses of raters' interview transcripts are also presented.

Chapter 7 summarises and synthesises the study's findings and evaluates them in light of topic validity requirements of a test. The implications of the findings are then elaborated on. The chapter concludes with a consideration of the study's limitations and suggests directions for future research.

Chapter 2 Literature Review

The aim of this chapter is to critically review different strands of scholarly research relevant to this study. Firstly, a survey of empirical research investigating the effects of topic, topic familiarity and/or background knowledge of topic on second language performance is presented within both task-based learning and performance assessment contexts. Secondly, findings from studies directly related to the IELTS speaking test and germane to the current study are discussed. Thirdly, the various features involved in performance assessment are presented followed by an introduction to the Rasch measurement model as a viable alternative to classical test theory in addressing the complexities of performance assessment. Fourthly, an overview of the current understandings of validity and validation is presented. Lastly, a description of Weir's (2005) socio-cognitive framework of language test validation and the way it is applied to the current study is presented.

2.1 Topic and Background Knowledge of Topic: Impact on Performance

2.1.1 Topic as a Task Condition: Impact on Task Difficulty

Topic, as a task-processing condition which can contribute to relative task difficulty, has been examined within a second language acquisition (SLA) task-based language learning and instruction research framework, most strongly associated with the works of Skehan (1996) and Skehan and Foster (1997). A description of this work is first provided.

Drawing predominantly on cognitive processing approaches to language learning, Skehan's (1996:52) framework provides a systematic and principled method in classifying tasks for pedagogic purposes. The three factors of *code complexity*, *cognitive complexity* and *communicative stress* are identified in the framework as factors affecting foreign language performance and contributing to task difficulty. *Code complexity* refers to the complexity of the language of task input in terms of lexical and syntactic difficulty as well as range. *Cognitive complexity* refers both to the processing demands of the task and to the familiarity of task content. Lastly, *communicative stress* refers to factors which might pose additional pressure on communication; these include time pressure, modality (e.g. speaking versus writing), scale (e.g. number of interlocutors), stakes (in terms of consequences of correct or incorrect task completion) and control (the extent to which learners can influence task achievement). These factors can be useful for categorising, sequencing and comparing different task types.

In the above framework, topic familiarity features in the cognitive complexity component of this framework and is associated with the *conceptualisation* stage of Levelt's (1989) model associated with 'generating an idea or set of ideas for expression' (Field, 2011:74) and 'accessing relevant aspects of schematic knowledge' (Skehan, 1996:52). In a later work, Skehan (1998) hypothesises that topic familiarity and drawing on information available as prior knowledge will have a positive influence on fluency and accuracy in oral performance. He argues that 'easy access to information should make only limited demands on attention, allowing material to be assembled for speech more easily, and with greater attention to form' (Skehan, 2001:175).

Skehan et al. (2012) also propose another general framework for task-based research in which studies can be categorised depending on whether they focus on pre-, during, or post-task use. Planning time, for example, is linked to the pre-task phase of the framework and has been widely addressed in research studies (Wigglesworth, 1997; Weir et al., 2006; Ortega, 1999; Mehnert, 1998). This is not surprising, as planning time is an objective variable which can be easily and precisely manipulated. Research on during-task choices can be more varied as the task (its difficulty and characteristics), as well as the conditions under which it is done, can be examined (Skehan et al., 2012). The post-task stage refers to the potential influence on performance of anticipating a post-task activity which is linked to the task.

From a theoretical perspective, topic familiarity in the Skehan et al. (2012:173) study is conceptualised as a form of *preparedness*, linked with the pre-task phase of the framework. A distinction is made between three forms of preparedness: preparedness brought about by planning time, preparedness owing to some form of task repetition or engagement with task-related materials (examples given are retelling a story or speaking about a topic after having written about it) and finally preparedness with the domain content:

Related, but distinct, is a second sense of preparedness: familiarity with the content domain involved. In other words, there may be areas of experience that are very familiar, or events that have occurred many times, or domains that have been studied more formally. In each case, the ideas that are to be expressed will not need excessive conceptualization, since they are available, perhaps as schemas, in long-term memory. In such case, the preparation has already taken place through the participant's previous life.

Skehan et al. (2012: 173)

Topic familiarity appears to be amongst the least addressed components of the Skehan's (1996) model within task-based performance research. Bei (2010) and Skehan et al.

(2012) draw on an earlier Skehan and Foster (1999) study – which examined the influence of task structure and processing load on performance in a narrative-retelling task – as an example of a study where the facilitative role of topic familiarity on fluency is evidenced. A closer look at the article, however, displays a rather loose link between structured/unstructured tasks and topic familiarity; task structure in Skehan and Foster (1999:100) broadly refers to structure in terms of the time sequence and predictability inherent in the task, hypothesised to potentially reduce the processing demands of tasks and thus enhance performance. To test the hypothesis, participants in the study watched two 8-minute episodes of the British television series *Mr. Bean* which were selected as representing two distinct degrees of sequential structuring in their respective story lines. In the ‘restaurant’ episode, Mr. Bean goes to a restaurant, orders food and then spends the rest of the time trying to hide the food. This episode is considered to display a ‘predictable basic sequence to the narrative’ (*Ibid.*: 103). In the ‘golf’ episode, however, Mr. Bean plays Crazy Golf and following a bad shot which takes him all across the city attempts to hit the ball back to the golf course without touching it so as not to break the game’s rules. This episode is considered as representing a relatively unstructured narrative with no predictable sequence. Findings suggested that the structured task enhanced fluency. Bei (2010) interpreted the structured nature of the ‘restaurant’ task as more schematically familiar and thus explaining participants’ gains in fluency. The findings of this study as evidence of the facilitative role of topic familiarity should be regarded with some caution, given that the extent to which the concepts of sequential task structure and topic familiarity can be causally linked was not made explicitly clear.

Further evidence for the positive influence of topic familiarity on performance comes from Bei's (2010) own study which examined the role of task preparedness (operationalised as topic familiarity) and strategic planning on spoken performance at different proficiency levels (measured using C-tests). A total of 80 Chinese undergraduates (40 computer science and 40 medicine majors) participated in this study. In terms of planning time, participants were evenly divided into a 10-minute planning group and a no-planning group. Bei's (2010) approach to topic selection was to opt for two topics that were as comparable as possible yet for which learners possessed significantly different levels of familiarity so that a 'clear-cut' familiar/unfamiliar distinction could be made. A preference was expressed for imposing 'hard', objective criteria in topic selection (2010:65), operationalised as matching/mismatching of topics to participants' academic disciplines; a 'natural virus' topic for the medicine majors and a 'computer virus' topic for the computer science majors constituted the matching condition. In addition, participants were asked to complete a topic familiarity questionnaire so that participants exhibiting familiarity with both topics could be removed. The two tasks are reproduced below from Bei (2010:66):

Topic 1: Please describe in detail the general process of the infection of virus in a human body, the possible consequences, and the general procedure for dealing with a virus-infected person.

Topic 2: Please describe in detail the general process of the infection of virus in a computer, the possible consequences, and the general procedure for dealing with a virus-infected computer.

Participants in the different planning conditions were then asked to speak on both topics so that their oral performance could be compared across the matched/mismatched topic familiarity conditions and across the two planning conditions.

The results of the study suggested significant effects for both familiarity and planning time in enhancing fluency, operationalised as speech rate and average mid-clause pausing. The relative effect sizes, nevertheless, indicated a more important role for planning time (a medium effect size) compared to topic familiarity (a small effect size). Given the length of the planning time (10-minutes) and the note-taking option (although notes were later removed), it is not unforeseeable that some of the planning time was dedicated to the rehearsal of content which could have subsequently affected fluency. Topic familiarity was found to have no effect on complexity although it did significantly affect accuracy, lexical sophistication and lexical diversity. In comparing performance across different proficiency levels Bei (2010:V) holds that ‘proficiency seemed to be concerned with forms rather than meaning expression, as higher proficiency participants always scored higher in accuracy and sometimes in complexity, but not so much in fluency or lexis’.

The above studies suggest that topic familiarity can have a positive impact on performance in relation to inter-language measures of fluency, complexity and accuracy.

2.1.2 Task Difficulty and Language Testing

The notion of task *difficulty* as a function of the cognitive demands and features of the task can be quite appealing from a language testing perspective, as it can hypothetically allow for the construction of a range of tasks from easy to difficult which could in turn distinguish between persons from a range of ability levels. Building on Skehan’s (1998) framework, Norris et al. (1998) put this idea to practice by devising a matrix of task features that could be systematically manipulated in order to develop prototype task-based performance tests that could be sequenced in terms of difficulty. In other words, by

locating different tasks within the matrix, *a priori* measures of task difficulty could be estimated.

Norris et al. (2002) then implemented the matrix by constructing 13 complex tasks which varied in difficulty as a function of different combinations of cognitive factors and subsequently evaluated task-based performance. Results were not as anticipated as the *a priori* difficulty estimates of the tasks failed to systematically predict observed differences in examinees' performances. Interestingly, these inconsistencies were traced back to examinees' familiarity with the tasks, both in terms of content and procedures.

The replicability of the SLA approach to determining task difficulty (as discussed in section 2.1.1) within a language testing (LT) context has been questioned by Fulcher (2003) who draws on evidence from several LT studies which failed to find significant influences on performance as a function of task difficulty. As way of explanation, Fulcher (2003:64) directs attention to the lack of '*score sensitivity*' of performance to changes in tasks: while the results of SLA studies (documented in the works of Tarone, 1988;1998) largely support that variations in tasks and task conditions (e.g. topic, number of interlocutors) may influence the discourse produced by test takers and different interlanguage measures, Fulcher (2003) and Fulcher and Márquez-Reiter (2003:326) challenge the 'unstated assumption that changes in discourse automatically translate into changes in test score, and hence, the estimate of task difficulty'. In support of this argument, the authors refer to the results of LT studies (e.g. Fulcher's 1993 comparative study of a picture description, text-based interview and group discussion tasks; Bachman, Lynch and Mason's 1995 comparative study of two different speaking task types) where substantial differences in task types only resulted in very small albeit significant

differences in scores and task difficulty measures. On the basis of these findings, the authors make a logical inference that if differences in task types fail to result in large effects on performance then by the same token, if task type is held constant, then changes in task conditions are unlikely to yield significant results with large effect sizes, and differences would only account for small variance in scores. The importance of establishing ‘practical significance’ is then highlighted by Fulcher (2003:65).

The term *practical significance* refers to differences which are not only statistically significant but are also associated with large effect sizes (see Kirk, 1996 for an excellent critical review of classical null hypothesis significance testing in which he also proposes the notion of practical significance). Dorans and Feigenbaum (1994) advanced the term ‘difference that matters’ or DTM in relation to SAT scores where ‘any differences less than the DTM are considered not big enough to warrant any concern since they are smaller than the smallest difference that might actually matter’ (Dorans and Liu, 2009:13). It is such large effect sizes and practically significant differences which are of relevance in language testing and high-stakes assessment contexts, as they can determine whether a candidate passes or fails a test and/or achieves a higher or lower speaking proficiency level.

In relating the above discussion to the present study, it becomes clear that by positioning the study within a language assessment context, there are two clear directions for its foci: first, that it is the extent to which topic and background knowledge of topic can have an influence on test *scores* rather than interlanguage measures which should be problematised; second, that it is the extent to which statistical significance translates into meaningful, *practical significance* that is of main concern. These two criteria should also

be borne in mind in reviewing language testing studies which have focused on the effects of topic and background knowledge of topic on performance. Before moving on to a description of these studies, it is important to complete the discussion on task difficulty by raising another fundamental problem with the concept.

2.1.3 Problematising Task Difficulty

Bachman (2002), in a well-reasoned argument, challenges the conceptualisation of *difficulty* as an inherent feature of the task making an explicit reference to the matrix in Norris et al., (1998) as exemplifying such *a priori* approaches to estimating task difficulty. Instead, he regards difficulty as an ‘artifact’ of test performance, reflecting the interaction between characteristics of the tasks and test takers’ language abilities:

The conceptualization of ‘difficulty features’ confounds task characteristics with test-takers’ language ability and introduces a hypothetical ‘difficulty’ factor as a determinant of test performance. In current measurement models, ‘difficulty’ is essentially an artifact of test performance, and not a characteristic of assessment tasks themselves. Because of these problems, current approaches to using task characteristics alone to predict difficulty are unlikely to yield consistent or meaningful results.

Bachman (2002: 453)

While Bachman (2002:466) acknowledges code complexity as a ‘unique’ characteristic of tasks, he views the other two components i.e. cognitive complexity and communicative stress as factors which interact with characteristics of the test taker and thus are not inherent to the task. He poses the question: ‘wherein lies difficulty?’ and illustrates the complexity involved in answering this question as follows:

Candidates, who will differ in their underlying competencies and ability for use, may find tasks with different qualities and conditions differentially difficult to perform. Different candidates will find different examiners and other interactants differentially easy or difficult to interact with. Different raters may apply the scale criteria differently to different performances, so that they may be differentially lenient or severe. So where is ‘difficulty’? I would argue that difficulty is not a separate factor at all, but ‘resides’ in the interactions among all of these components involved in assessment.

Bachman (2002: 466)

The above quote captures some of the complexities involved in performance assessment. It also highlights the need for a methodology that can adequately address these different concerns. In referring to the methodological implications of his argument, Bachman (2002) proposes that researchers identify and distinguish between (a) those task features which are inherent to the task and which do not require assumptions to be made about test takers and how they would potentially interact with the task; (b) characteristics of the test takers c) interactions between (a) and (b) and to subsequently conceptualise and model interactions as such.

These methodological considerations have direct implications for the proposed study's design but can also serve as a criterion in evaluating the methodologies adopted by other studies. In particular, whether BK of topic has been assumed in relation to specific topics or whether it has been established by other means.

2.1.4 Topic and Background Knowledge of Topic: Influence on L2 Performance

We now return our attention to a critical review of studies which have examined the role of topic and background knowledge of topic on L2 performance. The studies are classified according to the language skill they pertain to, that is, the receptive skills of reading and listening and the productive skills writing and speaking. The influence of the two variables on test-taker perceptions is subsequently considered. It should be noted that the purpose of this review is twofold; on the one hand, to establish whether the importance ascribed to topic and background knowledge of topic in the theoretical literature (Bachman and Palmer, 1996; Skehan, 1998; Weir, 2005) is reflected in the findings from empirical research and on the other hand, to critically reflect on the methodologies adopted in

various studies in informing and shaping the design of the proposed study. For the latter reason, the review also includes studies from language learning contexts. However, the majority of the studies are drawn from language testing literature.

2.1.4.1 Reading

The results of empirical research on the effects of prior knowledge on reading comprehension have been inconsistent and far from conclusive. The majority of this research concentrates on general vs. academic topics with some studies reporting a clear background/prior knowledge effect (Alderson and Urquhart, 1983; Chen and Donin, 1997; Krekeler, 2006) and others reporting a lack of a consistent effect on performance, frequently citing language proficiency as a possible confounding factor (Alderson and Urquhart, 1985; Clapham, 1996).

In a series of studies by Alderson and Urquhart (1983; 1985), the effects of background knowledge on the reading comprehension of international students in pre-sessional courses at university were investigated. Background knowledge was not independently measured but was assumed on the basis of students' past or future subject area – an arguably flawed assumption as pointed out by Davies (2008). The findings were perplexing in that the first study showed that students performed better when tested in their own subject area (findings echoed in Chen and Donin, 1997) whereas the subsequent study revealed inconsistencies across disciplines. For example, economics students performed better than the engineering students on the economics texts but the opposite advantage was not found. The authors suggested the likely effect of a proficiency threshold level above which background knowledge exerts more influence on performance, but they did not provide further evidence in support of this assertion.

The most frequently cited evidence in support of a *lack* of a systematic prior knowledge effect comes from the seminal work of Clapham (1996). In her study, the relative influences of background knowledge and language proficiency on reading comprehension were investigated. 842 participants each took two versions of a reading test: one relevant to their field of study and one from a different field. The reading passages in the 10 specific reading subtests were analysed in terms of specificity by bringing forward perspectives from students, content area experts and applied linguists and it was found that the texts varied substantially in their degree of subject specificity from general to highly specific. Background knowledge was established using a questionnaire of reading habits and familiarity with content area.

Echoing Alderson and Urquhart (1985), findings from the Clapham (1996) study revealed inconsistencies in terms of a background knowledge effect, as some students performed better on tests in their own field of study compared to other fields, but this was only the case when the passages were highly subject-specific. The trend was not observed when the passages were more general (e.g. passages extracted from introduction to academic articles vs. passages describing research processes in specific fields). While significant variance was explained by both language proficiency and background knowledge, it was also found that background knowledge differentially affected students at different proficiency levels and that it was the medium proficiency group of students who were most influenced by their background knowledge.

In light of these findings, the explanation put forward by Clapham is as follows:

While lower level students could not take advantage of their background knowledge because they were too concerned with bottom-up skills such as decoding the text, and while high proficiency students were able to make maximum use of their linguistic skills so that, like native speakers, they did not have to rely so heavily on their background knowledge, the scores of medium proficiency students were affected by their background knowledge.

Clapham (2000:515-516)

On the basis of these results she concluded that subject-specific tests may not be 'equally valid' for learners at different English ability levels and recommended the abandonment of English for Specific Academic Purposes; a recommendation which informed the revision of IELTS in eliminating subject-specific reading and writing models for different academic disciplines (Clapham, 2000).

The hypothesis that students may draw on their background knowledge above a certain proficiency threshold (Alderson and Urquhart, 1985) was put to the test by Krekeler (2006) in a study of background knowledge on reading comprehension of 486 international students at German universities in a languages for specific academic purposes (LSAP) assessment context. In addressing the weaknesses of previous studies in measuring background knowledge, Krekeler used a different approach to evaluating background knowledge by a) asking test takers to explain key terminology in the texts before test administration (only two terms per text); b) asking test takers with a binary yes/no question whether they had previous exposure to the text topic and c) taking the test takers' future subject field into account. Language proficiency was established using a C-test. While these measures are an improvement to the many assumption-based measures of background knowledge, they are arguably too crude a measure in capturing nuances in background knowledge. Nevertheless, these measures, particularly the binary yes/no question can be a clear-cut indication of familiarity. Findings from the study showed

highly significant differences between means on the reading test scores of the two groups according to all three background knowledge grouping criteria. In terms of an interaction between background knowledge and proficiency level, the background knowledge effect was shown to be equally important for both high and low language proficiency students with the exception of very advanced test takers. There was very little evidence to substantiate the threshold hypothesis on the basis of the analyses from the interaction effects, leading the author to suggest that ‘if thresholds exist at all, they are fuzzy and may even be chance events’ (Krekeler, 2006:123). The author then emphasises the important role of background knowledge in LSAP tests in that ‘it would be more sensible to assume that background knowledge, or lack thereof, will usually affect test performance’ (Krekeler, 2006:123) but cautions that the effect may not always be predictable.

A comprehensive and methodologically sound design was adopted by Usó-Juan (2006) in an attempt to overcome the limitations of other studies on the effects of discipline-related background knowledge on reading comprehension in an English for Academic Purposes (EAP) context. 380 Spanish L1 undergraduates took part in the study. Six reading passages from three disciplines (two passages per discipline) of psychology, tourism and industrial engineering were selected. Subject specialists were consulted for text selection, and readability statistics were calculated for the purposes of text comparability. A range of reading comprehension question types was also included (e.g. matching items, true-false items, open-ended questions, summary task, etc.). Prior to the reading tests, participants’ level of English and level of background knowledge for each text were measured using proficiency and topic knowledge tests respectively. Multiple-choice questions were used as a measure of topic knowledge and included ten items, which either tested explicit

information from the passage or pertained to more general knowledge for each topic. Results of the multiple regression analyses revealed significant influences of both variables with approximately 21-31% of the EAP reading scores accounted for by discipline-related background knowledge and language proficiency explaining 58-68% of the variance. Using delineated regression equations, the author also examined the relative compensatory effect of the two variables and arrived at the following conclusion:

Successful EAP reading is possible without discipline-related knowledge if the participants' English proficiency level is advanced or intermediate. However, if the participants have a low level of proficiency in English, successful EAP reading is possible if the participants reach a linguistic threshold and have discipline-related knowledge.

Usó-Juan (2006: 222)

Moving away from the English for academic and specific purposes domains, the positive role of culture-specific background knowledge in enhancing comprehension, recall of information and strategies used by readers have been documented in several studies (Johnson 1982; Malik, 1990; Steffensen, Joag-dev and Anderson, 1979; Carrell, 1981). For example, when the advanced Japanese and Chinese participants in Carrell's (1981) study were exposed to English translations of folk tales from their own cultures as well as other Western European and American-Indian cultures, results indicated that participants read, understood and recalled information from texts that were closely related to their own familiar culture better than texts which dealt with less familiar cultures.

A methodological concern was raised by Carrell and Wise (1998) in relation to studies which subsume prior knowledge and interest in topic under one construct. Their research was set up to separate these two variables in a study of 104 students of ESL in an American EAP programme. The criteria behind text selection were to include topics that were of potentially varying levels of both interest and prior knowledge for the students so

that each student could be administered with texts in four conditions of high/low background knowledge and high/low topic interest. To this end, ten topics were selected from encyclopedia entries which covered a range of topics from computers, human evolution, Islamic art and architecture to Olympics and stock exchange. Prior knowledge of topics was elicited with a multiple-choice factual knowledge test and topic interest was elicited using self-reports in a rank-order scale from 1-10. Proficiency level was not independently measured and instead, programme-level grouping was used as indication of language ability. Reading comprehension was measured using ten multiple-choice questions comprising of both textually explicit and implicit questions. Once students had completed the prior knowledge test and topic interest form, a combination of passages that best fitted the four conditions was tailor-made for each individual in the study and subsequently administered. The correlation coefficients between prior knowledge scores and topic interest rankings demonstrated that the two did not always correlate strongly. The results of the repeated-measures analysis of variance (ANOVA) showed that L2 proficiency was the only variable with a significant main effect. While the main effects for prior knowledge and interest did not reach significance, the differences for the two low and high groups were in the expected direction. On the other hand, significant first-order interactions were observed for prior knowledge and topic interest. It was shown that if both prior knowledge and topic are low, there is a significantly lower mean score for reading comprehension whereas higher prior knowledge and topic interest enhances reading comprehension to a small extent. The authors conclude that 'prior knowledge and topic interest are not highly significant nor additive factors in reading comprehension' (Carrell and Wise, 1998:299).

In sum, the results from the above studies remain inconclusive. In their review of the literature, Clapham (2000) and Krekeler (2006) refer to several factors which can explain these inconsistencies, the most salient of which has been the role that proficiency level plays in potentially influencing the way in which BK operates in performances. Other factors include degree of text specificity, topic and the different operationalisations of background knowledge.

2.1.4.2 Listening

The results of studies on the effects of prior knowledge on listening test performance are also mixed. Schmidt-Rinehart (1994), for example, found topic familiarity to be a ‘powerful factor at all levels of proficiency’ (185). In contrast, Jensen and Hansen (1995) failed to find reliable topic effects on listening comprehension performance.

Schmidt-Rinehart’s (1994) study aimed to examine the role of topic familiarity and proficiency in a project with 90 university students of Spanish. Listening passage familiarity was determined *a priori* and on the basis of previous exposure to the information in the course syllabus and was later substantiated with a follow-up familiarity questionnaire. Similar to Carrell and Wise (1998), proficiency level was established according to students’ course levels. This is surprising, given that the study explicitly aimed to examine proficiency as a factor and thus, course level might be considered an unreliable measure of language ability. The design involved students listening to a familiar passage (Hispanic universities) and a novel passage (Going for a Walk in the Park) although given the title of the latter, the extent to which ‘novelty’ translates into unfamiliarity can be called into question. The post-listening questionnaire was later used to exclude students who indicated familiarity with the novel passage. Listening

comprehension was measured using an immediate written recall task in the participants' L1 (English). The results of a repeated-measures ANOVA showed significant main effects for both familiarity and course level with no significant interactions between the two leading the author to conclude that 'background knowledge in the form of topic familiarity emerges as a powerful factor in facilitating listening comprehension' (Schmidt-Rinehart, 1994:183). However, the author acknowledges that the lack of interaction effect could be attributed to lack of variation in participants' proficiency levels.

Conflicting results were found in Jensen and Hansen's (1995) study of the effects of prior knowledge in understanding lectures. In testing the hypotheses that listening proficiency moderates the effects of prior knowledge on lecture comprehension and that prior knowledge is a strong predictor of lecture comprehension, data from 11 lecture subtests of a Test of Listening for Academic Purposes (T-LAP) were analysed for this study. Similar to one of Krekeler's (2006) measures, prior knowledge of topic was measured with a binary yes/no question at the end of the test, the results of which were used to categorise students into two groups: prior knowledge and no prior knowledge. The scores from the non-academic section of the T-LAP were used as a measure of listening proficiency. Findings from multiple regression analyses revealed significant interaction effects for only one of the 11 lectures (on History). As expected, a significant main effect was found for listening skills for all lectures. However, a significant main effect was found for prior knowledge in only five of the 11 lectures with a small effect size, accounting for 3-9% of the variance.

Topic familiarity, listening proficiency and discourse modification (in terms of adding information redundancies and elaborations in listening text passages) were the three

variables explored in relation to listening comprehension in academic discourse contexts in Chiang and Dunkel's (1992) study with 360¹⁰ male-only Chinese EFL listeners. Listening ability was measured with the listening section of the Comprehensive English Language Test on the basis of which students were divided into two groups of low and high listening proficiency. Selected lectures were on a culturally familiar topic (Confucianism) and a culturally unfamiliar topic (Amish people) respectively. The modified versions of each passage included elaborated information in the form of paraphrase and information repetition. Students in each proficiency level were then randomly assigned to one of the four combinations of familiar/unfamiliar and modified/unmodified of passage input. Listening comprehension was measured with an information-recognition quiz consisting of both passage-dependent and passage-independent items with the latter serving as an indication of participants' prior knowledge. Results pertaining to prior knowledge indicated a significant interaction between prior knowledge and test type. Participants achieved markedly higher scores on the familiar topic (Confucianism) but only on passage-independent items. Regardless of topic, no significant effect was found for prior knowledge on performance in passage-dependent items. Earlier, it was mentioned that the authors included the passage-independent items as a measure of prior knowledge. However, in interpreting the data the authors state:

If a listener has a schema of the lecture content, as had the Chinese listeners who listened to the lecture on Confucius, the listener will be able to process the information and achieve better comprehension of the lecture. In other words, the more prior knowledge the listener has about the topic of the lecture, the easier it is for that listener to comprehend the lecture and retain general points of information.

Chiang and Dunkel (1992:365)

¹⁰ The study originally had 388 participants but the scores from twenty eight students who displayed 'intermediate' level listening scores were removed from the analyses.

The extent to which the study provides evidence in support of the facilitative effect of prior knowledge in listening comprehension can be challenged. The higher scores on passage-independent items solely confirm the authors' original assumption that Confucianism was indeed more familiar to students. In other words, the passage-independent items serve as a topic knowledge test. If passage-independent items were administered before the lecture as a base-line measure of prior knowledge and were once again administered after the lecture and there was evidence of significant improvement in scores, then the authors would be justified in their interpretation. However, as it stands, the interpretation is speculative and the evidence from the study does not adequately lend support to a positive effect of prior knowledge on listening comprehension.

The use of topics which draw on markedly different levels of background knowledge is a common method in examining the role of topic familiarity in more experimental contexts, although its relevance to language test settings where differences in topics are far less pronounced can be questioned. For example, the effects of religious-specific background knowledge on listening comprehension scores were studied by Markham and Latham (1987) in texts of prayer rituals. On the basis of self-reports, the 65 participants in the study were divided into three groups (Muslims, Christians and religion-neutral) and subsequently listened to passages associated with prayer rituals in Islam and Christianity. Findings suggested a clear trend of increase in mean scores when there was a match between passage content and religious background specifically in the recall of idea units. The findings from the religion-neutral group did not display a consistent pattern.

A positive influence of topic-related schemata was found in Long (1990) who examined the influence of background knowledge and linguistic knowledge on the listening

comprehension of 188 students of Spanish as a foreign language. A repeated-measures design was used where participants' listening comprehension was first tested on a topic hypothesised to be unfamiliar (gold rush in Ecuador) and then on a topic hypothesised to be familiar (U2 rock band). Participants' linguistic knowledge was not directly measured but was established through grades received in the previous Spanish course as well as self-assessments of listening performance. Prior to hearing the texts, a general background knowledge survey was administered 'as a means of probing for and activating' (Long, 1990:70) the schemata relevant to the two texts. After hearing each text, participants were asked to summarise the information they heard in their L1. They also completed a checklist consisting of paraphrased statements in English in relation to the passage content but interspersed with distractors. They were instructed to only check those statements mentioned in the passage. Listening comprehension was subsequently measured on the basis of total number of idea units, the proportion of correct idea units as well as scores on the checklist. Findings from the survey results confirmed the researcher's hypothesis in that participants' background knowledge was significantly lower on the gold rush topic compared to the U2 topic. Participants also scored significantly higher on the number of idea units and proportion of correct idea units. However, no significant differences were found for the checklist results. These results lend support for the positive influence of prior knowledge. However, the author also draws attention to a subset of 13 recall protocols where there was evidence of 'dysfunctional effects' of schemata on listening comprehension across both topics. Examples included mixing of temporal details or extending information from the probe survey to the listening passages which led the author to err on the side of caution in drawing conclusions regarding the positive influence of topic familiarity stating that 'schemata can hurt, as well as help' (Long, 1990: 72).

2.1.4.3 Writing

A clear trend for the positive role of background knowledge on performance was found in Tedick (1990) where the written performance of 105 English as a Second Language (ESL) graduate students was reported to be significantly better – across all proficiency levels – on a topic related to the participants' field of study compared to a general topic. Performances were rated holistically on a six-point rating scale. A closer examination of the study reveals that the field-specific prompt in the study provided the writers with some flexibility in choice, as the writers were instructed to select a controversial issue in their field of study, to discuss the controversy and to take a position in relation to it. It was reasonably argued that by providing this choice, participants were more likely to draw on a topic with which they are familiar and which can thus serve as a determinant of background knowledge. The study's findings were taken to suggest that 'writing performance improves qualitatively when the subjects are familiar with the subject matter of the writing stimulus' and when the topic 'allows them to make use of their prior knowledge' (Tedick, 1990:132). In looking at the differences between the two prompts, Lim (2009:37) questions field-specificity of the prompt, arguing that the specific prompt is 'ironically (...) the more general prompt' as it is 'virtually unconstrained, leaving respondents plenty of leeway on what to write about'.

In a similar vein to Tedick (1990), Jennings et al. (1999) conceptualised test takers' 'choice of topic' as an indication of their background knowledge in examining the topic validity of an integrated test of reading, listening and writing. In defending their methodology, the authors point to the complexities of interactions between a given topic and test takers' prior knowledge of the topic, interest in topic and perceptions regarding

the relevance of the topics and argue that when test takers make a choice, the choice is a reflection of the relative salience of these features for the test taker. Unlike Tedick, (1990), the findings from this study failed to substantiate a positive role for BK.

The 254 ESL university applicants in the Jennings et al. (1999) study were randomly assigned to *choice* (of five topics) and *no-choice* conditions and performances were compared across the two conditions on four dependent variables of overall proficiency level, reading response, lecture response and essay response. Test-taker perceptions of the effects of choice were also elicited. A strength of the study was to control for task comparability by selecting an *a priori* topic for the no-choice group and subsequently separating from the choice-group those candidates who selected the same *a priori* topic from the five available topics. However, due to this same design, there was no control over the number of participants selecting the *a priori* topics which is why the number of participants in the two groups were not equivalent. Results of the Mann Whitney U test revealed no significant differences between the two groups for any of the dependent variables. Nonetheless, the mean scores for the choice group were found to be consistently higher than the no-choice group and the most marked difference was found in the essay task.

The non-significant findings are not unexpected, as the integrated nature of the test designed to equip test takers with the prior knowledge necessary to complete the tasks, potentially attenuates the effects of topic. These results were taken as validity evidence for the test holding that the ‘context provided by the test materials reduces the impact of prior knowledge to the point of insignificance’ (Jennings et al., 1999: 448). From the test takers’ standpoint, however, the test topics were found to be the second most important

factor influencing performance after time allowed for the test. . The authors draw attention to this apparent mismatch between the perceived significance of test topic impact on performance and actual test results and recommend a choice mechanism as a ‘means of alleviating’ some test takers’ concerns with topic effects (*Ibid.*: 449). The positive impact of choice of topic on test taker attitudes is also documented in Kenyon and Malabonga (2001) who ascribed an important role to choice in assessment contexts in promoting test-taker autonomy on the one hand, and in providing opportunities for test takers to demonstrate their proficiency levels, on the other.

The topic validity of the writing section of an ESL placement test where topics were randomly assigned to test takers was also the focus of a large-scale study by Lee and Anderson (2007). The study was motivated by the need for empirical evidence which demonstrated the comparability of the academic topics in the study and their generality across different groups of test takers. The impact of the three variables of general language proficiency, academic topics and background knowledge of topics on performance scores were examined using multi-nomial logistic regression analysis. Background knowledge was not explicitly measured but was assumed on the basis of the students’ departmental affiliation and candidates were subsequently categorised into humanities, technology, life sciences and business. It was hypothesised that ‘topics generated from specific disciplines are more difficult for students majoring in fields distant from the topic area’ (Lee and Anderson, 2007: 318). The results of the study indicated that once proficiency was controlled for, there was a main topic effect, which suggested that given a specific ability, the probability of achieving a given score would be dependent on the assigned topic and would thus undermine topic comparability. In contrast, no interaction was found between

candidates' departmental affiliation and topic which the authors drew on to conclude that the writing topics were general enough to be used for students with varying degrees of background knowledge. One can argue that departmental affiliation is too broad a categorisation to be used as an indicator of background knowledge or topic familiarity, as it is primarily assumption-based. The authors acknowledge that an independent measure of background knowledge which is more 'critically related' to the topics could have greatly enhanced the findings of the study. Interestingly, there is little reference to the integrated nature of the writing test (with a video-taped lecture and a reading article) as an explanatory factor for the findings, given that such tests can equip the test taker with the background knowledge necessary to respond to prompts and thus reduce its effect on performance (Jennings et al., 1999; Weigle, 2004; Huang, 2010).

In a study which focuses primarily on general topics in a standardised assessment context, He and Shi (2012) raise the question, from a fairness standpoint, of whether the topics of prompts used in such tests are 'general enough not to require particular cultural or subject-specific knowledge' (444). On the basis of a pilot study, the authors identified the two prompts of 'university studies' and 'federal politics' as topics associated with 'general' and 'specific' topical knowledge respectively. These prompts were administered to a group of 55 students from three proficiency levels (based on the results of a placement test) and performances were compared in terms of analytic scores on content, organisation and language. The findings from the study revealed a clear pattern of significantly higher scores on the general knowledge task compared to the more specific task on the composite overall score, with significant main effects for proficiency level, prompt type as well as a significant prompt-proficiency level interaction. The same trend was observed across the

component scores. However, no effect sizes were reported. Other quantitative findings suggested that the difference between the two prompts was most salient for the advanced level participants and that ‘idea quality’ and ‘position taking’ under the content component were the two features most markedly influenced by the different prompts. Results of qualitative analyses identified the challenges associated with generating enough ideas for the specific prompt, the culture-dependent nature of the topic and the lack of topic-specific vocabulary as the most salient themes emerging from the post-test interviews.

On the basis of these findings the authors challenge the notion that language proficiency is the main determinant of writing performance and suggest the inclusion of prior knowledge of topic as a significant factor in writing. They further recommend a distinction to be made between ‘L2 writing ability from L2 writers’ interpretations of writing topics resulting from their prior knowledge and L1 cultural backgrounds’ (He and Shi, 2012: 460). While the findings from this study are illustrative in showing a clear, positive advantage for topic knowledge in performance assessments, it can also be argued that the design of the study was set up to come up with significant differences; a point also acknowledged by the authors who subsequently recommended the use of ‘less obviously oppositional topic choices’ (*Ibid.*: 461) for further research on the subject.

2.1.4.4 Speaking

Empirical research on the influence of topic and background knowledge of topic in speaking performance contexts are quite limited and similar to the other skills, the findings are equally equivocal.

The appropriateness of using general test topics to select ESL International Teaching Assistants (ITAs) to teach their subjects in English-speaking classrooms in the US was questioned by Smith (1989;1992) who hypothesised that performance may be enhanced on tasks designed to reflect ITAs' specific academic field compared to more general topics. To address the hypothesis, Smith designed three field-specific versions of the SPEAK test (for chemistry, physics and mathematics) and subsequently compared the scores of ITAs on the general version of the test with scores from the field-specific versions on features of fluency, pronunciation, grammar and overall comprehensibility. While the ITAs' background knowledge of the field-specific topics can be assumed, their level of background knowledge of the general topics was not independently measured or considered as an explanatory factor. Results of the study did not reveal a consistent trend with some ITAs performing better on the field-specific tests and others on the general version with very small, non-significant differences in mean values on features of comprehensibility, pronunciation and fluency. Grammar was the feature where no difference was observed between the mean scores, leading the author to conclude that 'insofar as these two tests are able to characterize it, the factor of topic alone does not bring about significant group differences in the oral performance of second language users' (Smith, 1989:161). However, when looking at the pass/fail cut off scores, these small differences in scores could have made the difference between passing or failing for some of the ITAs in the study. The latter finding is a reflection of Fulcher's (2003) notion of 'practical significance' (Fulcher, 2003) discussed earlier (see section 2.1.2).

Negative influence of field-specific prior knowledge was found in Douglas and Selinker's (1992) comparative study of a field-specific (CHEMSPEAK) and a general (SPEAK) test

of oral proficiency which was designed to evaluate which of the two better predicted the results from a third teaching performance test (TEACH). CHEMSPEAK was modelled on the SPEAK test in terms of format. However, modifications were made to the test in terms of language and instructions so that the test would approximate an academic setting. 31 prospective ITAs took all three versions of the test during the course of a day and their performances on the CHEMSPEAK and SPEAK tests were scored by 15 and 12 raters respectively. Scores were averaged across raters. Rather unexpectedly and counter to the researchers' hypothesis, the lowest mean comprehensibility scores were observed on CHEMSPEAK compared to the other two tests with significantly lower scores on CHEMSPEAK compared to SPEAK. In spite of these findings, the CHEMSPEAK scores were found to correlate significantly and invariably with the TEACH raters' recommendations for ITA teaching. Put differently, the CHEMSPEAK scores were a better predictor of raters' final decisions on whether or not an ITA should be allowed to teach, in comparison to SPEAK scores. The authors proposed three rationalisations for these contradictory findings. Firstly, that the participants were focused more on content than on form when discussing chemistry. Secondly, that the tasks were more complex in CHEMSPEAK. Thirdly, that the raters displayed a greater degree of conservativeness as a function of unfamiliarity with the chemistry-specific content. In conclusion, the authors state that 'at least for this group, it is *not* the case that 'if they [ITAs] were tested in speaking about their own fields' they would automatically and universally do better' (Douglas and Selinker, 1990: 323). This is one of the only studies which ascribes a negative influence to background knowledge. It is also the first study discussed so far which briefly alludes to the potential role of rater characteristics in influencing performance assessment; a subject which will be discussed in section 2.2.

Rather than distinguishing between field-specific vs. general topics, the focus of Papajohn's (1999) systematic and comprehensive study is on the comparability of academic topics. Limited to the field of chemistry, the oral performances of 102 prospective ITAs on 15 different field-specific chemistry topics in the chemistry TEACH test were investigated. Background knowledge of topics was assumed for all chemistry graduate ITAs, given that the topics covered basic chemistry. It should be noted that linguistic performance, and not content, was the subject of assessment and that ratings were given by language teachers and not chemistry professionals. The three criteria of concepts, math and calculations were used for topic classification and were operationalised as the cognitive demand they pose on the topic in terms of relative abstractedness, the level of math required for problem solving and the number of steps required for problem solving respectively (Papajohn, 1999). It was hypothesised that 'the more complex the concepts, math and calculations of a given topic, the more difficult it will be for an examinee to convey the information clearly to raters who may have limited knowledge of chemistry' (*Ibid.*:62). The results of a multiple regression analysis suggested that both general language proficiency (using scores from the SPEAK test) and topic groupings were significant predictors of the TEACH scores accounting for 67.2% and 4.7% of the variance respectively.

The strength of the Papajohn (1999) study lies in the systematic way in which topics were analysed and classified, the methods for which can be useful for other academic fields but perhaps not equally suitable or relevant for more general topics and tests. Attention should be drawn to two important factors which may have confounded the study's findings: teaching skills and rater familiarity with topics. The test instructions required test takers to

explain their assigned topic as if they were teaching it to a class of undergraduates. Their performance was subsequently rated on pronunciation, grammar and fluency within an overall comprehensibility category. However, as Papajohn (1999:76) argues ‘it might be difficult for raters to separate language ability from teaching ability’. An examination of the interaction between topic and rater familiarity with the topic could have also greatly enhanced the study in providing ‘useful data on the relationship of shared background and comprehensibility’ (Papajohn, 1999:78). Bachman’s (2002) argument regarding ‘difficulty’ as a test artifact rings particularly true in this example where it is possible to question whether the difficulty of topics are an inherent feature of the tasks or whether, for example, they partly reflect raters’ familiarity with some of the topics.

Inconsistent topic effects were reported in Lumley and O’Sullivan’s (2005) study of large-scale tape-mediated exit test of English in Hong Kong in which the effects of task topic, test-taker gender and audience on the spoken performance of 894 university students were examined. Task topics were classified as neutral, male-oriented or female-oriented. Two versions of each topic were also developed where the intended hypothetical audience was manipulated to be either male or female. Using Multi-faceted Rasch Measurement (MFRM), the effects of each variable and their interaction on performance scores – assigned using an analytic rating scale on the four criteria of task fulfilment and relevance, clarity of presentation, pronunciation, and grammar and vocabulary – were examined. The results of a gender by task difficulty bias analysis suggested a significant yet small bias for a small subset of tasks (four tasks out of 27) where the direction of the bias was split with two tasks favouring males and two favouring females. Further exploration of the data revealed that for one of the more distinctly male-oriented tasks (entertainment and horse

rating), there was a small advantage for males on the task fulfilment and relevance criterion with a female audience but that there was a more pronounced difference with a male audience on three of the criteria and particularly, the task fulfilment and relevance criterion. The general picture that emerged was that ‘topic is more significant than audience, while an interaction of the two compounds the effect’ (Lumley and O’Sullivan, 2005:430). Where there was evidence of bias, it was reported to be small and in the most part limited to one criterion and not always stable across different test forms leading the authors to suggest that ‘the effect is insufficiently reliable to indicate systematic bias’ (Lumley and O’Sullivan, 2005:431). It can also be argued that the classification of topics on the basis of assumptions i.e. according to preconceived notions of what is stereotypically male or female may have had confounded the results. As acknowledged by the authors and in line with Bachman (2002), the authors hold that ‘task difficulty is too complex to be categorized in such simplistic terms as stereotypical notions of what is interesting or familiar to men and women. Tasks are more likely to affect individuals differentially, rather than at the level of groups defined at this level’ (Lumley and O’Sullivan, 2005:432-433).

A strong positive influence for topical knowledge on performance in integrated and independent test tasks was found in Huang (2010). The study adopted structural equation modeling and qualitative analysis of interview data to examine the role of topical knowledge and test anxiety on spoken performance of 352 Taiwanese students of EFL. This is one of the only studies which adopted a rigorous approach to measuring topic knowledge by constructing and extensively piloting a series of topic knowledge tests corresponding to the task topics used in the study. These tests were developed as a

prediction measure where test takers were asked to predict whether specific ideas would appear in a hypothetical text on specified topics. These predictions were then compared against judgements by experts and served as a measure of degree of background knowledge. One possible weakness of such an approach is that it might be easy to guess which ideas are relevant to a given topic. Also, the fact that these tests were administered prior to the speaking tests could have potentially primed the test takers regarding the kind of topics they were expected to respond to.

Amongst Huang's (2010) main findings was that topical knowledge had a strong positive influence on integrated speaking performance and as expected, anxiety had a strong negative influence. Topical knowledge was found not to have a direct impact on anxiety although, given the low-stakes test performance setting of the study, this finding is not surprising (Huang, 2010). It was also found that the topic of tasks exerted an influence on the way topical knowledge impacted on performance across independent and integrated speaking tasks. This study could have greatly benefited from the inclusion of a general language proficiency measure in modeling performance.

2.1.4.5 Summary and Methodological Implications

The previous sections presented a critical review of studies which have focused on the effects of topic and background knowledge of topic on performance. The findings of these studies in respect to the relationship between topics, background knowledge of topics and performance have been shown to be inconsistent and largely inconclusive. These inconsistencies might be explained by the myriad ways in which variables such as background knowledge or performance have been operationalised. For example, as classified by Huang (2010), background knowledge has either been *assumed* on the basis

of factors such as gender, cultural background, current or future academic field of study or has been *inferred* on the basis of a variety of methods including self-report questionnaires, previous experience with the topics/texts, reading habits, academic field of study and topic knowledge tests. The literature suggests a priority of inference-based methods of assessment over assumption-based methods due to the highly individual nature of background knowledge which does not lend itself easily to assumptions. This has important methodological implications for the current study in terms of measuring background knowledge.

The review of the studies also suggested that background knowledge might differentially affect performances of test takers from different proficiency levels. However, findings from different studies revealed inconsistencies which might, once again, be partially due to the different ways in which language proficiency has been measured in various studies (e.g. from self-reports and classroom groupings to results on independent tests of proficiency such as C-tests). It is therefore important for the current study to include a reliable and independent measure of general language proficiency in examining topic and BK effects on performance.

Lastly, the review of the literature holds a further important methodological implication. As recommended by Bachman (2002), in exploring the influence of the variables of interest on performance scores, it is necessary to adopt a method which allows for the isolation of topic (as a task characteristic) from background knowledge of topic (as a test-taker characteristic).

2.1.5 Rater and Test-Taker Perspectives: Research from IELTS

The results of studies focusing on rater and test-taker perceptions and experiences of test features can be informative in pinpointing problematic areas and in guiding research efforts. For this reason, results of studies directly focusing on the IELTS speaking test and which bear close relevance to the current study are considered in this section. These include a survey of IELTS examiners (Brown and Taylor, 2006), a small-scale IELTS washback study (Smith, 2009) and a qualitative inquiry into the rating process of IELTS examiners (Brown, 2006). Only the findings pertaining to topic and topic-related features are reported here.

A concern with a potential topic effect in speaking tests was one of the strongest emerging themes in a large-scale worldwide survey of 269 IELTS Examiners (Brown and Taylor, 2006). This study was motivated by a major revision to the IELTS Speaking test in 2001 and was designed to elicit examiners' views and experiences in relation to the modifications in the new test (e.g. the introduction of an examiner script and the substitution of a holistic scale with an analytic scale).

Regarding topics, findings from examiners' general comments pointed to the cultural irrelevance and inappropriateness of some topics for candidates from particular parts of the world. Two questionnaire statements elicited examiners' views on the a) suitability of the topics for the candidature and b) the equivalence of topics in terms of difficulty. These statements were repeated for each of the three parts of the Speaking test. Topics were found to be suitable by 55%, 70% and 74% of the examiners for Parts 1, 2 and 3 of the test respectively (see section 1.6.2 for a description of each part). Part 1 topics were found to be least suitable. In terms of topic equivalence, the results are telling and far less positive.

The majority of the examiners did not find the topics to be equivalent in terms of difficulty (66%, 61% and 52% corresponding to topics of Parts 1, 2 and 3 respectively). Examiner comments touched on several issues regarding topics. Firstly, the incomparability of topics in terms of difficulty, appropriateness and/or complexity; secondly, the unsuitability of topics as a function of age, culture and level of proficiency of candidates. Some topics were described as ‘dull, simplistic, too abstract or too obscure’ (Brown and Taylor, 2006:16).

These findings suggest that tasks which are designed to be parallel may be perceived as exhibiting differing levels of difficulty as a function of their topics or the interaction of topics with test-taker characteristics. IELTS research on whether such perceptions match statistical data for tasks/topics is limited and/or not made publicly available by the IELTS partners. In fact, the paucity of quantitative empirical research on the comparability of tasks and test forms in large-scale standardised assessment tests has been highlighted by Weir (2005).

The examiners in the Brown and Taylor (2006) study also referred to topics in relation to the speaking assessment criteria (emphasis is given to these findings, as they relate to the rating scale that is used as an instrument in the study and therefore have important bearings for the study’s methodology). Firstly, some examiner remarks referred to the inadequacy of criteria in capturing the performances of candidates ‘who might be fluent but speaking off-topic’ or ‘say very little’ (Brown and Taylor:16-17). To address this limitation, the researchers suggested the inclusion of a ‘task-response scale’ or ‘a specific focus on how the candidates address the questions’. To the best of my knowledge, these recommendations have not been implemented in the scales. Secondly, in respect to the

Fluency and Coherence scale, examiners reported difficulties in distinguishing whether the sources of hesitation in candidate speech were content or language related (Brown and Taylor, 2006:16-17). The latter finding strongly resonates with the results of Brown's (2006) qualitative inquiry into the validity of the IELTS rating scales.

Brown (2006) analysed the verbal reports of six IELTS examiners as they rated previously recorded speaking tests using a stimulated recall technique in order to identify specific features of the rating scale that examiners struggled with or found problematic when making rating decisions. One of the strongest themes in the study was the challenge for examiners to discern or infer the reasons behind hesitations, repetitions or pauses in candidate speech in relation to the *Fluency and Coherence* scale:

They [examiners] frequently attempted to infer the cause of hesitation, at times attributing it to linguistic limitations (...) and at times to non-linguistic causes_ to candidates thinking about the content of their response, to their personality, to their cultural background, or to lack of interest in the topic (having "nothing to say"). Often examiners were unsure whether language or content was the cause of disfluency but, because it was relevant to the ratings decision (...) they struggled to decide. In fact, this struggle appeared to a major problem as it was commented on several times, both in the verbal reports and in the responses to questionnaires.

Brown (2006:48-49)

These findings suggest that topic-related factors in test-taker performances might have an effect on rater decision-making when assigning scores.

Test-taker concerns regarding the choice of topics in the IELTS Speaking Test emerged as a salient theme in a small-scale study of IELTS washback. Smith (2009) used questionnaires and focus groups to explore the exam preparation practices of postgraduate students at the University of Oxford. The study also attempted to elicit the participants' perceptions of the Academic IELTS test and features of the test that they found to be particularly salient as well as the subsequent influence of the test on their academic lives. It is not surprising to observe that topic, once again, is a recurring issue in this study.

Reflecting on their test day experience, some of the participants referred to factors such as ‘overall luck, topic luck and the examiner’ as contributing to their scores. These factors were categorised as ‘external factors’ by the researcher drawing on Weiner’s (1992) *attribution theory* (Smith, 2009:51-52). Moreover, ‘topic difficulty’ turned out to be a particularly problematic issue for half of the focus group participants (7 out of 14). Topic-related problems included having little to say about the topic (even in their L1), having little interest in the topic, finding the topic ‘mundane and silly’, being unable to relate to the topic and experiencing anxiety as a result of topic unfamiliarity (Smith, 2009:59-60). A quote by one of the participants in reference to the speaking test illustrates some of these difficulties:

I remember that it was something completely unrelated to me so actually I had to made up all the, all the information I was giving him... I had no real information to, in which to build my answer so I just invented a whole story.... Actually (...) that made it a little bit difficult because I didn't, I just didn't have like to think, like the English, I wanted to use for that I also to create information and then, express it properly in English which took some time of my mental [...] processing.

Participant I4 in Smith (2009:59)

The participant’s observation in linking the added cognitive processing demand of the task when topic was unfamiliar resonates with the cognitive complexity component of Skehan’s (1996) task difficulty framework.

2.2 The Rater Factor: Impact on Performance

So far, we have examined the effects of two related factors on performance in a second language: topic, as a characteristic of the task and background knowledge of topic, as a characteristic of the test taker. In performance assessment, the influence of a third factor i.e. the rater – as inextricably linked to the assessment process and as a source of variability – cannot be understated. As McNamara points out (1996:177): ‘performance

assessment necessarily involves subjective judgements. This is appropriate: evaluation of any complex human performance can hardly be done automatically'¹¹.

The pivotal role of raters lies in the link that they create between the test takers' performance, the rating scale and the scores. However, by involving human judges in the assessment process, a degree of subjectivity or an 'Achilles heel' (O'Sullivan and Rignall, 2007:447) is simultaneously introduced to performance assessment. Issues subsequently arise from the interdependence of such subjectivity with measurement error (Cronbach, 1990), most critically associated with relative rater harshness and leniency, score unreliability, the introduction of construct-irrelevant variance to the test and the potentially unintended consequences for test takers as a result of rater-related classification errors (McNamara, 1996; Lumley and McNamara; 1995; Bachman, Lynch and Mason, 1995; Wiseman, 2012).

A frequently cited quote by Edgeworth (1890:653 in Linacre, 1989:10) reveals the rater effect as a long-established phenomenon, contributing to substantial variation in candidate scores:

I find the element of chance in these public examinations to be such that only a fraction - from a third to two-thirds - of the successful candidates can be regarded as safe, above the danger of coming out unsuccessfully if a different set of equally competent judges had happened to be appointed.

Notice how the 'chance' element i.e. which judge happens to be doing the judging is distinguished from the 'competence' of the judges. It appears that it is not the judges' ability or expertise, which is in question, but rather how their individual differences

¹¹ Since 1996 various advancements in the field of speech recognition and speech processing technology (Van Moere, 2010) have made the automated scoring of speaking tests possible (e.g. Pearson's PTE academic speaking test, the TOEFL iBT speaking practice tests). A discussion of these tests and the automated approach to scoring is out of the scope of this thesis.

(seemingly in terms of harshness and leniency) can contribute to large variations in candidate scores. This observation, from well over a century ago, reflects findings from recent studies into the effects of rater training – aimed presumably at enhancing competence and reducing variability – which have shown partial improvements in reliability of scores (Shohamy, Gordon, and Kraemer, 1992) but with little success in reducing rater harshness/leniency tendencies (Weigle, 1998).

Shohamy, Gordon, and Kraemer (1992), for example, examined the effects of raters' background (lay vs. professional) and rater training (training vs. no training) on score reliability in a writing assessment context. Findings revealed that regardless of training or background, raters had demonstrated an ability to achieve acceptable levels of reliability in scoring. However, extensive training was also shown to improve reliability although (any) differences in rater severity and leniency were not commented on.

On the other hand, Weigle's (1998) systematic analysis of score data from 60 essays using Multi-Faceted Rasch Measurement (MFRM), aimed to distinguish between the effects of training on a) consistency in rating and b) relative harshness/leniency of both experienced and inexperienced raters. Findings suggested improvements in rater self-consistency as a function of training. However, substantial differences in severity persisted which led the author to ascribe a positive role to training in 'helping raters give more predictable scores (i.e. intra-rater reliability)' as opposed to 'getting them to give identical scores (i.e. inter-rater reliability)' (Weigle, 1998: 263). A similar conclusion was reached in Lumley and McNamara's (1995) longitudinal study of ratings awarded by 13 experienced raters after two rater training sessions with an 18-month gap in between and a third session following an operational test administration two months after the second training session. Training

was found to be ineffectual in narrowing the differences in raters' harshness and leniency and had 'by no means eliminated, nor even reduced [variation] to a level which should permit reporting of raw scores for candidate performance' (Lumley and McNamara, 1995: 69). On the basis of these findings and the consequences of such substantial variation on candidate scores despite training, the authors call into question the adoption of a single-marking approach by exam boards. Instead, they recommend the use of double/triple markings or preferably, MFRM 'since it is able to take relative severity of judges into account and make adjustments to estimates of candidate ability' (*Ibid.*:69).

While Lumley and McNamara's (1995) study was limited to experienced raters, Lim's (2011) large scale study of novice and experienced raters – also in a written performance assessment context – investigated raters' scoring behavior, in terms of consistency and severity, over three time periods. Relevant to this discussion is the finding that lack of experience in rating is not necessarily associated with a particular pattern of severity:

Where severity is concerned, it appears that novice raters may or may not be distinguishable from experienced raters. There were new raters who were much more severe or lenient compared to the average, but there were also new raters who performed similarly to existing raters from the moment they began marking.

Lim (2011:551)

A more systematic trend was observed in Lumley and O'Sullivan (2000 cited in Taylor and Galaczi, 2011:210) where newly certified raters were shown to exhibit more severity and consistency whereas experienced raters exhibited a tendency towards leniency and at times, inconsistency.

A different approach to decreasing examiner variation was adopted in O'Sullivan and Rignall (2007). Following Wigglesworth's (1993) observation that feedback to raters may enhance rating quality, O'Sullivan and Rignall (2007) explored the usefulness of

providing formal feedback (from the results of MFRM bias analyses) to trained IELTS raters in improving rating quality in terms of severity and consistency. Once again, and in line with previous findings, the gains from this ‘one-shot feedback’ (O’Sullivan and Rignall, 2007: 469) were found to be limited. The results of questionnaire feedback from raters, on the other hand, revealed a positive and motivating impact of the feedback on raters’ decision-making processes. Drawing on their results, where even experienced and trained examiners exhibited instances of bias, the authors question the value and functionality of training from an inter-rater agreement perspective and instead recommend training efforts to be concentrated more on intra-rater agreements (O’Sullivan and Rignall, 2007).

Three important issues are now considered in light of the consistency and stability of the above findings. The first is a question which turns the problem of rater variation on its head i.e. ‘is perfect agreement even desirable?’ (McNamara, 1996:126). The second issue is a consideration of *why* training efforts have been unsuccessful in reducing rater variability in assessment contexts. The third is the issue of *how* to deal with such variation. These issues are first discussed before briefly touching on how MFRM can provide a theoretical and practical solution to some of these issues.

The allure of perfect rater agreement in ensuring reliability was problematised by Constable and Andrich (1984) who cautioned that such reliability might be achieved at the risk of validity:

It is usually required to have two or more raters who are trained to agree on independent ratings of the same performance. It is suggested that such a requirement may produce a paradox of attenuation associated with item analysis, in which too high a correlation between items, while enhancing reliability, decreases validity.

McNamara (1996: 126) draws on the notion of the *Rashomon* effect to question whether there is one definitive judgement of performance which is ‘true’. The term refers to the subjectivity of perception and borrows its name from a film title by one of the masters of classic Japanese cinema, Akira Kurosawa in which a crime witnessed by four observers is described in four ‘contradictory yet equivocal’ ways (Fanselow, 1977:17). The term ‘has come to embody a general cultural notion of the relativity of truth’ (Kamir, 2000:41). Bringing the discussion back to assessment, McNamara (1996) argues that the interpretations of raters from the same performance, while different, may be equally valid and that trying to find a ‘definitive’ judgement is futile. Instead, he suggests accepting rater variations as ‘a fact of life’, directing training efforts towards increasing raters’ internal consistency and to aim to compensate for such differences through multiple marking or more sophisticated approaches to measurement such as MFRM. In MFRM, raters are conceptualised and subsequently modeled not as ‘scoring machines’ but as ‘independent experts’ (Linacre, 2011a) where the individual differences between raters are not viewed as hindrances to be eliminated but rather as expected features of the rating process which can be modeled and subsequently accounted for (Weigle, 1998).

The second issue pertained to the reasons why training appears to be ineffective in reducing rater subjectivity. One explanation lies in the complexity and cognitive demand of the rating process (Cronbach, 1990; Wiseman, 2012). Other explanations refer to the multitude of systematic rater tendencies or ‘rater effects’ which might be traced back, for example, to the ways raters interpret and understand a scale or the performance features that they find salient (Pollitt and Murray, 1996; Eckes, 2010) and can subsequently manifest themselves in a number of ways, the most prominent of which is differences in

severity. Myford and Wolfe (2003; 2004) and Linacre (2011a) identify the most common rater effects as follows:

- The severity/leniency effect, which refers to raters' relative likelihood of assigning higher and lower scores (Wiseman, 2012) either compared to other raters or to a benchmark (Taylor and Galaczi, 2011)
- The halo effect, which can be described as a rater's tendency to award similar ratings than justified on different criteria of an analytic rating scale (Marais and Andrich, 2011)
- Extremism/central tendency effect, which refers to a rater's tendency to award scores in either the extremes or middle categories of a rating scale
- Playing it safe effect (Linacre, 2011a), which refers to a rater's disposition towards assigning scores which are similar to those awarded by other raters
- Interaction or bias effect, which refers to 'a pattern of harshness/leniency with regard to one or more aspects of the rating process' (Taylor and Galaczi, 2011: 209) such as particular test takers, tasks or criteria (Eckes, 2010).

Distinct from systematic rater effects, is the 'random effect' which refers to idiosyncratic and unpredictable patterns of rater behaviour which cannot be explained in a systematic fashion (McNamara, 1996).

This section has demonstrated the various measurement errors which can be introduced to a performance assessment context as a consequence of systematic and random rater-related effects. Whether in operational or experimental test settings, the potential impact of the rater on performance scores has to be taken into account and addressed.

2.3 Performance Assessment and Multi-Faceted Rasch

Measurement

Following McNamara's (1996) model of test performance (Figure 1-1, p.2), the discussions so far have illustrated how different aspects of a testing situation, henceforth referred to as *facets*, for example, task, test-taker and rater characteristics, are at play in contributing to a score observation. The question which arises is how to isolate the relative influence of these facets from one another? In other words how is it possible to determine whether a given score reflects the actual ability of the test taker, the difficulty of the task, the severity of a rater who marked the task or a combination of all these factors? These are fundamental psychometric issues which have been raised in relation to classical sample-dependent approaches to measurement and which point to the unreliability of raw scores as an indication of the underlying ability of the test taker.

Bachman summarises some of the above drawbacks as follows:

There are many situations, particularly in large-scale assessment, for which classical item analysis is inadequate because the item statistics we obtain are dependent on the particular sample of test takers we try them out with, while test scores are dependent on the particular set of items we give to test takers. This makes it difficult to compare items across different groups of test takers, and to compare test takers across different tests. Thus, in situations where we want to use our tests with different groups of test takers, or want to develop multiple forms of the test that yield parallel results, we need a more powerful way than item analysis to obtain statistical information about test items. Item response theory (IRT) can provide such a means.

Bachman (2004:152)

Bachman (2004:152) then lists the three most common Item Response Theory (IRT) models in modern test theory: the one-parameter or Rasch model in which the difficulty of the item parameter is estimated, the two-parameter model, in which both the difficulty and discrimination are estimated and finally the three-parameter model, in which difficulty, discrimination and guessing are parameterised.

In the present study, it is the Rasch model and its extension to Multi-Faceted Rasch Measurement (Linacre, 1989) which has been adopted as an alternative approach to measurement in a performance assessment context, as it allows for various facets of the testing situation to be modeled simultaneously but examined independently. MFRM has now become a widely-practiced approach in ‘*rater-mediated*’ contexts which involve ‘constructed-response items’ and where scores are awarded using rating scales (Eckes, 2009).

Before explaining the model, attention should be drawn first to a significant source of controversy between those who see the Rasch model as a special case of IRT models (e.g. Bachman, 2004 above) and those who see the Rasch model as substantially different (see Andrich, 2004 for a comprehensive account of the origins of the controversy). Andrich (2004:I-7) makes a case that not only are these perspectives different, but that they are irreconcilable and reflect ‘incompatible paradigms’ in terms of the data-model relationship. The next section further elaborates on this controversy, as it is important to understand the paradigm to which the current study belongs.

2.3.1 Approaches to Measurement: a Case of Different Paradigms?

A *paradigm* (Kuhn, 1962) in research is a term that characterises a set of concepts, ideas, perspectives and approaches to research and reflects a common understanding and way of thinking about reality. In Andrich’s (2004:I-7) words, the term refers to ‘a collection of mutually reinforcing understandings underpinning a science’. It is not so much the details of how research is carried out in different paradigms – even within the same field of inquiry – that distinguishes one from the other but rather, the shared perceptions of and attitudes towards scientific research of the communities working within each paradigm.

In terms of the data-model relationship, the perspective that views the Rasch model as a special case of IRT classes of models will be hereby referred to as the traditional paradigm. In this approach, the aim of measurement is to construct a model that best fits the given data with no constraints on the class of models and parameters that can be used. If the one-parameter model does not fit the data, then the two-parameter model can be used and so on and so forth. In contrast, in what shall be referred to as the Rasch paradigm, there is an *a priori* constraint on the class of models and parameters that can be used, that is, only those models which are compatible with principles of fundamental measurement in the physical sciences. These models do not characterise data and are independent of data. Therefore, in this approach, it is not the model that needs to fit the data but rather, it is the data that need to fit the model (Andrich, 2004). In case of data misfit, rather than opting for another model, the Rasch paradigm invites a qualitative inquiry into the original data.

This study conforms to the Rasch paradigm. We now return our attention to a description of the Rasch model, its extension to MFRM and illustrate how the model can be applied in the proposed study.

2.3.2 What is Rasch Analysis?

Rasch analysis is the ‘formal testing of an outcome scale against a mathematical measurement model’ developed by the Danish mathematician Georg Rasch (Tennant and Conaghan, 2007: 1358). Rasch abstracted a class of ‘probabilistic models of measurement’ on the basis of empirical work on ‘intelligence and attainment tests’ (Rasch, 1960). The key characteristic of the model is that it operationalises the formal axioms that

underlie measurement (Luce and Tukey, 1964)¹² and this has very important implications for the social sciences as ‘the principles and properties of conjoint measurement...would bring the same sort of rigorous measurement to the human sciences as those in the physical sciences have enjoyed for a considerable time’ (Bond and Fox, 2007:14).

The model’s principle of ‘invariance’, put forward by Rasch, is as elegant as it is simple:

A person having a greater ability than another person should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another means that for any person the probability of solving the second item is the greater one.

Rasch (1960: 117)

Notice the use of the term *probability* in the above quote. The Rasch model, which is a probabilistic model, can be contrasted with deterministic models. The latter have little application in the social sciences, as data are rarely deterministic (Andrich, 2005). Conversely, the Rasch model uses a probabilistic counterpart of Guttman scaling¹³, which meets the criteria of fundamental measurement, to provide a framework against which data from the social sciences can be formally tested. (Andrich, 2005; Tennant and Conaghan, 2007). Simply put, the Rasch model can inform us of the extent to which data can form an ordinal or interval level scale.

¹² ‘The essential character of what is classically considered...the fundamental measurement of extensive quantities is described by an axiomatization for the comparison of effects of (or responses to) arbitrary *combinations* of "quantities" of a *single specified kind*...Measurement on a ratio scale follows from such axioms. ...the essential character of simultaneous *conjoint* measurement is described by an axiomatization for the comparison of effects of (or responses to) *pairs* formed from *two specified kinds* of "quantities."...Measurement on interval scales which have a common unit follows from these axioms; usually these scales can be converted in a natural way into ratio scales’ (Luce and Tukey, 1964:1).

¹³ Guttman scaling (Guttman, 1950) is a ‘deterministic pattern that expects a strict hierarchical ordering of items’ (Tennant and Conaghan, 2007:1358). It holds that the total score should predict exactly which items were answered correctly or incorrectly. It is viewed as an ideal, in providing evidence of the unidimensionality of a construct (Andrich, 1982).

It should be noted that one of the main underlying assumptions¹⁴ of the Rasch model is that of *unidimensionality* which is ‘a basic concept in scientific measurement that one attribute of an object ... be measured at a time. The Rasch model requires a single construct to be underlying the items that form a hierarchical continuum’ (Bond and Fox, 2007:314). The assumption holds that the items/ prompts/ tasks in a test or questionnaire are designed to measure the same unidimensional latent trait, and that the ability of the persons and difficulty of the items can be located along the same dimension. The more ‘familiar counterpart’ of unidimensionality in latent trait theory is the notion of internal consistency in traditional test theory (Andrich, 1982: 95), where items in a test are designed to reflect the same thing (Kuder and Richardson, 1937; Cronbach, 1951). When scores from items on a test are summed to calculate a total score or an average score, it is assumed that all the items are measuring the same trait.

Whether or not a construct is truly unidimensional can be challenged. However, as Andrich and Marais (2010:1) point out, it is misleading to raise this question as ‘at some level of precision, no construct is unidimensional, while at some levels of precision, any construct is’. Instead, they recommend evaluating this question in terms of the purpose for which the measurement is used.

In its simplest form, the dichotomous Rasch model for a unidimensional construct holds that the probability of person (v) giving a response (x) of 0 or 1 to a dichotomous item is a logistic function of the relative distance between the ability of a person (β_v) and the

¹⁴ The other main assumption of ‘response independence’ in the Rasch model will be covered in Chapter 4.

difficulty of an item (δ_i). If β_v is larger, equal to or smaller than δ_i , then we can expect the following probabilities¹⁵:

If $(\beta_v - \delta_i) < 0$ then $P\{x_{vi} = 1\} < 0.5$

If $(\beta_v - \delta_i) = 0$ then $P\{x_{vi} = 1\} = 0.5$

If $(\beta_v - \delta_i) > 0$ then $P\{x_{vi} = 1\} > 0.5$

We can therefore see that the probability of a given response is abstracted as the distance between the person's ability, and the item's difficulty. The Rasch model for a dichotomously scored item therefore takes following form:

$$P\{x_{vi} = 1 \mid \beta_v, \delta_i\} = \frac{e^{(\beta_v - \delta_i)}}{1 + e^{(\beta_v - \delta_i)}}$$

The formula can be read as the probability of person v getting a correct answer (1) to item i given the person's ability (β_v) and the item's difficulty (δ_i). This relationship can be graphically viewed in the item characteristic curve (ICC) for a dichotomously scored item (i) in Figure 2-1.

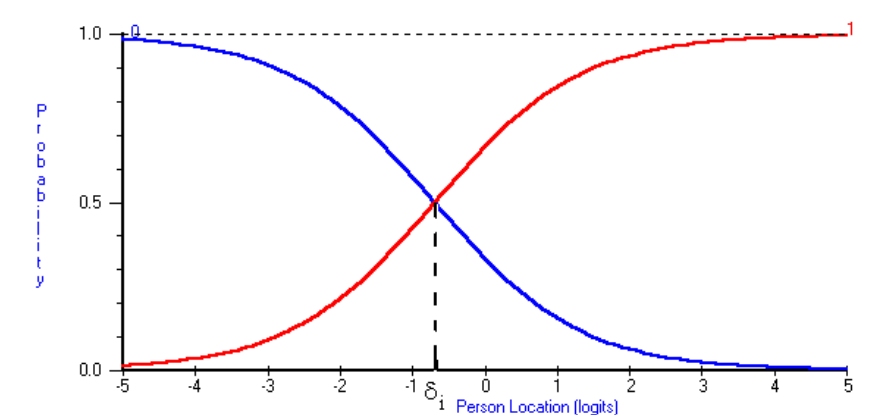


Figure 2-1 Item Characteristic Curves for a Dichotomously Scored Item (i)

¹⁵ The formulae in these sections are reproduced from Andrich and Marais (2010).

The horizontal line represents person ability in logits¹⁶ while the vertical line displays probability values. The blue curve represents the model's expectations for a score of 0 while the red curve represents expectations for a score of 1. As we move along the person ability continuum, or as ability increases, the probability of a getting a 0 decreases while the probability of getting a correct answer of 1 increases. The location in which the difficulty of the item is equal to the ability of the person (marked with vertical dotted line) is where there is a 50-50 chance of the person getting the item right or wrong. In the Rasch model, the abilities of the persons and the difficulties of the items are expressed on the same common frame of reference – an interval level logit scale – which allows for ability and difficulty to be meaningfully compared (McNamara, 1996).

The dichotomous Rasch model has been extended to rating scale (Andrich, 1978) and partial-credit (Masters, 1982) models, and has been applied to attitude questionnaires and polytomous items i.e. items with more than two response categories (see Chapter 5 for applications of these models to the instruments in this study).

Another extension of the Rasch model is the Many-Facet Rasch Model (Linacre, 1989; 1994), and is expressed as follows:

$$\log (P_{nikj}/P_{nikj}-1)= B_n - D_i - F_k - C_j$$

Where B_n is the ability of person n , D_i is the difficulty of an item, task or criterion i , F_k is the threshold of score k and C_j is the severity of judge j . In the Many-Facet Rasch Model (MFRM), the person parameter can be conditioned out and the rater and item (or prompt, task, criterion) parameters can be simultaneously estimated (Andrich and Marais, 2011).

¹⁶ Logit refers to the unit of measurement used in the Rasch model when the ordinal level data are transformed to log odd ratios to form an interval level measure scale. The mean of the items on a test are arbitrarily averaged at 0.00 logits in the Rasch model.

The ordinal relationship between all these parameters is then expressed on a common interval scale where facets of the assessment context can be meaningfully compared to person abilities.

MFRM can be extended to include more parameters. In applying the model to the current study, the facets of the assessment context include test-taker ability, topic difficulty, difficulty of background knowledge levels, rater severity and criteria difficulty. MFRM then allows for an examination of the influence of the main facets of interest i.e. topics and background knowledge of topics, independent of but adjusted for the influence of the other facets.

2.4 Topic Validity and Spoken Performance Assessment

Having reviewed the literature on the test facets of interest in the study and selected a psychometric model of performance assessment, the next step is to select a theoretical model and framework which provides a structured and systematic approach to gathering validity evidence in order to address the study's research question. First, an overview of the current understandings of validity and the validation process are presented. Second, Weir's (2005) socio-cognitive framework of language test validation is described. The framework is then applied to this study by guiding the formulation of subsidiary research questions, the answers to which can bring together different types of validity evidence for building the study's overall topic validity argument.

2.4.1 Validity

Validity lies at the heart of educational measurement and language assessment alike and is viewed as 'the most fundamental consideration in developing and evaluating tests' in the

current *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999:9). Language tests are measurement instruments, designed to tap into an underlying construct (e.g. communicative language ability) and serve as a link between the construct of interest and score data generated by the test. As Cizek (2011) highlights, it is the *quality* of the score data which is of critical concern to a given test's stakeholders, particularly in high-stakes settings where scores are used to inform decisions about individual test takers and therefore can have important consequences. A common example in language testing is when scores on a standardised language test are used as an admission criterion for higher education academic programmes.

Traditionally and in the early days of psychometrics, validity was defined as whether a 'test measures what it is supposed to measure' (Lado, 1961:321) and was believed to be an inherent attribute of a test. However, the 1980s saw a shift in the conceptualisation of validity, culminating in the seminal work of Samuel Messick (1989) who moved away from a view of validity as discrete types (e.g. content, criterion, construct validities) to a single, unified concept defined as follows:

Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness of inferences and actions* based on test scores or other modes of assessment (italics in original).

Messick (1989:13)

This -now classic- definition of validity¹⁷ captures the essence of current thinking in validity theory. There are several important aspects of the contemporary approaches to validity that are important to understand and they are summarised below on the basis of the literature (Messick, 1989; Linn and Gronlund, 2000; Bachman, 2004; Weir, 2005;

¹⁷ It should be noted that Messick's (1989) definition of validity has been criticised, from a semantic perspective, to 'conflate validity (a property) with validation (a process)' (Cizek, 2011:3).

Cizek, 2011). Specific references are made to Weir (2005) in order to demonstrate the extent to which his validation framework is in line with current understandings of validity.

1. Validity relates to interpretations and inferences based on test scores and not the test itself i.e. the centrality of the test has given way to the centrality of score-based inferences (Linn and Gronlund, 2000). This is echoed in Weir (2005:12) who holds that ‘validity is perhaps better defined as the extent to which a test can be shown to produce data i.e. test scores, which are an accurate representation of a candidate’s level of language knowledge or skills.’
2. The traditional view of discrete types of validity (e.g. criterion, content, concurrent, predictive) has now been replaced with a unified view of validity with construct validity at its heart (Bachman 2004). From this stance, what was previously considered as different types of validities are now viewed as different types of evidence which can be brought together collectively to lend support to the interpretations and inferences made about the construct that the test is claiming to measure (Cizek, 2011). The view of construct validity as a superordinate category comprising of different types of evidence is echoed in Weir (2005:13) who refers to the ‘multi-faceted’ nature of validity.
3. A validity argument consists of ‘several parts’, which, on their own, cannot adequately substantiate the inferences which are to be based upon them, but rather it is the coherent synthesis of these parts which can lead to a holistic judgement on the appropriateness of test score use (Bachman, 2004:260). Similarly, Weir (2005:13) emphasises the necessity for collecting multiple sources of evidence which are not seen as alternatives but as ‘complementary aspects of an evidential

basis for test interpretation’ where ‘deficit in any one raises questions as to the well-foundedness of any interpretation of test scores’.

4. Validity cannot be viewed as a dichotomy or in Messick’s own words ‘validity is a matter of degree, not all or none’ (1989:13). Validity is only as good as the logic of the arguments and evidence brought forward to support it and can thus never lend itself as ‘proof’ of a given ability (Linn and Gronlund, 2000). As Cizek observes, the fact that validity cannot be viewed as black or white follows naturally from the ‘integrated evaluative judgment’ element of Messick’s quote (1989:13) as ‘those who evaluate validity evidence arrive at subjective judgements about the extent to which the corpus of accumulated evidence supports the intended test score inferences’ (Cizek, 2011:5). Likewise, Weir (2005) describes validity as a ‘relative’ concept.
5. Validation is seen as an ongoing process: ‘just as it is incorrect to say that a *test* is valid, so it is incorrect to say that the validity case for an intended inference is closed’ (Cizek, 2011: 5).
6. A consideration of the social consequences of test use¹⁸ and the integration of evidence of test consequences in a validity argument is seen as integral to a unified concept of validity (Messick, 1989). Commonly referred to as consequential validity, the term can be understood as ‘the functional worth of scores in terms of social consequences and their use’ (Messick, 1989: 84). The primacy of consequential validity evidence is emphasised by Stobart (2008:2) who states that

¹⁸ Prior to Messick (1989), concerns with the social consequences of test use were also raised by language testers. For example, Madsen (1983) used the term ‘affect’ in relation to test-driven anxiety; a consideration of the rationale behind testing or the ‘ethics’ of testing, was put forward by Canale (1987) and finally Hughes (1989:1) defined ‘backwash’ as ‘the effect of testing on teaching and learning’ and later expanded it to ‘the impact of tests on learners and teachers, on educational systems in general and on society at large’ (Hughes, 2003:53).

‘it is what you do with a test that determines validity’. The concept has also been embraced by proponents of *critical language testing* (Shohamy, 2001) who, drawing on the work of Foucault (1975) and from a social critical theory standpoint, view tests as instruments of power. Broadly speaking, there appears to be a general consensus amongst researchers and practitioners that test consequences are important. However, in a recent, superbly argued article by Cizek (2011), the *place* of consequences as *part of validity theory* has been heavily challenged. What lies at the crux of Cizek’s (2011) argument is that the ‘meaning, interpretation or inference of the test result, that is the *validity* of the test result, is unaffected by the actions that follow its use, whether positive or negative’ and therefore, the justification for test use ‘neither supports nor weakens the case for the validity of score inferences’ (2011:16). In other words ‘once evidence regarding the intended score interpretation has been gathered, the validity – that is, the meaning, of scores obtained from administration of the instrument – is not changed depending on the particular use to which those scores are put (Cizek, 2011: 17). In his view, the validation of inferences – a predominantly technical and scientific process – should be carried out independently yet parallel to an evaluation of potential test consequences which falls in the domain of ethics. To this end, he proposes an alternative heuristic where a conceptual distinction and a practical one are made between the validity of test scores on the one hand and the justification for test use on the other (see Cizek, 2011).

The above discussion sheds light on some of the common understandings as well as controversies in validity theory. In the next section, a brief overview of the process of

validation is presented. The two terms have not always been distinguished clearly in the literature (Cizek, 2011) but as O’Sullivan and Weir (2011:19) highlight, ‘the distinction is important as we see the model of validity as the underlying theory and the model of validation as the manifestation or operationalisation of this theory’.

2.4.2 Validation

The process of validation is ultimately about collecting, synthesising and evaluating evidence in constructing a supporting base for the plausibility of the argument about the interpretation of a given test score (Kane, 1992; Cizek, 2011). Sources of evidence can be ‘theory, logical argument and empirical evidence’ (Cizek, 2011:26).

One of the critical issues raised about validity and the validation process is that they have largely remained in the domain of theory and have yet to be operationalised and applied in a systematic fashion. Cizek (2011:12) for example, views a failure on behalf of validity theorists in ‘delineating’ the boundaries of validation efforts. As a result, practitioners may be faced with an open-ended and ongoing research agenda where there is no clear indication of a beginning or an end. This concern is echoed by Kane (2001) who criticises the unsatisfactory nature of validation efforts for lack of clear criteria:

A [validation] program invites such an eclectic and possibly unending process, it is not clear that the program does much to discourage an opportunistic strategy based on readily available data rather than more relevant but less accessible evidence. If an essentially infinite number of studies are relevant, where should one start, and how much is enough?

Kane (2001:327)

A useful guideline and starting point in directing validation efforts, offered by both Kane (2001) and Cizek (2011), is to prioritise those types of evidence that can shed light on two fundamental sources of validity threat: (a) construct underrepresentation (or misspecification) which can occur if important features of the construct that can directly

influence performance have not been elicited and/or measured by the test (Cizek, 2011) and (b) construct-irrelevant variation which refers to the influence on performance of variables other than the one intended to be measured by the test (Kane, 2001). According to Kane (2001:327) such validation efforts can ‘restrain empirical opportunism’. One of the foci of the study reflects the latter of the two approaches i.e. a consideration of the construct-irrelevant factor of background knowledge of topic in affecting performance.

2.4.3 Weir’s Socio-Cognitive Framework of Language Test Validation

The previous section presented a brief overview of the current understandings (and controversies) of the contemporary theories of validity. For the current study, Weir’s (2005) socio-cognitive framework of language test validation has been adopted for guiding the study’s validation efforts on the grounds that it represents a practical framework of language performance assessment *and* validation. To illustrate, the Weir (2005) framework a) allows for an examination of language performance in terms of the individuals’ underlying mental abilities (a cognitive perspective), contextual features (a social perspective) as well as the interactions between them while also including an evaluative parameter b) captures the current thinking in validity theory c) lends itself easily to operationalisation and correspondingly provides guidance on possible methodologies for collecting and synthesising validity evidence. These features are in line with McNamara’s (1996) criteria for an adequate model.

Weir’s (2005) framework (later modified in O’Sullivan and Weir, 2011) consists of six main components (see Figure 2-2): *test taker characteristics*, *cognitive validity*, *context validity*, *scoring validity*, *consequential validity* and *criterion-related validity*. The conceptual relationship between these different components is shown via the arrows,

which represent ‘the principal direction(s) of any hypothesised relationships’ (Weir, 2005:43). The model also has a temporal element which is used to indicate what kind of validity evidence needs to be collected and *when*. According to Weir (2005:43), the time sequence ‘runs from top to bottom: before the test is finalized, then administered and finally what happens after the test event’.

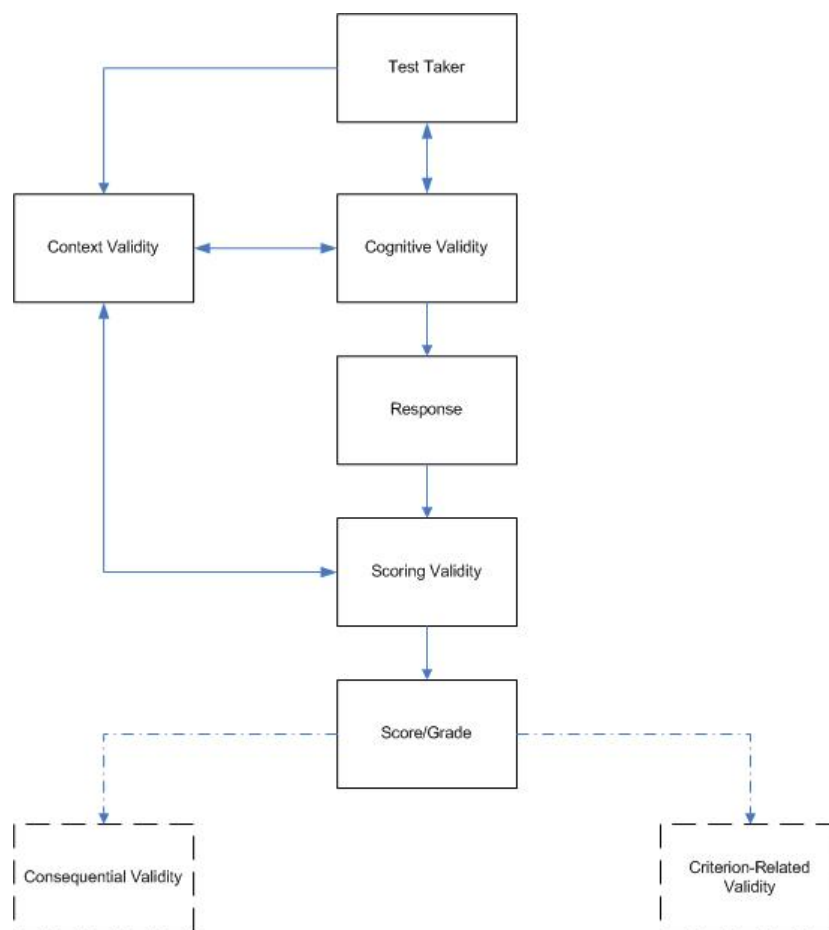


Figure 2-2 ‘A Socio-Cognitive Framework for Test Development and Validation’ (O’Sullivan and Weir, 2011:21)

As its socio-cognitive title suggests, in the framework, an ‘approximation’ to the underlying construct is conceptualised as the interaction between both cognitive elements

and contextual features (Weir, 2005). This is visually depicted in the bi-directional arrows between the context validity and the cognitive validity components of the model. Moreover, the model depicts the centrality of the test taker (and his/her characteristics) as exerting a direct influence on the way the contextual parameters of the model are cognitively processed (Weir, 2005). Once again, the arrows represent the hypothetical interactions between characteristics of the test taker, characteristics of the task and the mental processes which have been activated by the task parameters all of which lead to the production of a response. The response is subsequently scored (e.g. through interaction with rating scales, raters, etc.) and again, we see the hypothetical interactions between the scoring validity component and both context and cognitive validity components. These contextual (social), cognitive and evaluative (scoring) components constitute the ‘core’ elements of the framework (O’Sullivan and Weir, 2011:24) and even though they are visually presented as being independent, Weir (2005) emphasises that there is a ‘symbiotic’ relationship between the elements of the framework which, taken together, depict his conceptualisation of unified approach to validity.

Messick’s (1989) concern for a consideration of test consequences is reflected in the consequential validity component of the model which, along with criterion-related validity, is located at the bottom of the graph and *after* a score has been generated. Temporally speaking, these two types of validity evidence can be generated once a test has been developed and evidence of construct validity has been established. Conceptually speaking, these elements are viewed as ‘external’ to construct validity and not as core components, a position which resonates to a certain extent with Cizek’s (2011) stance on

separating the ‘core’ and technical process of construct validation from the justifications of use.

The framework has conceptualised validity as ‘multifaceted’ where different sources of validity evidence (context, cognitive, scoring, consequential and criterion-related) are fitted together to present a unified approach to validity. The model is therefore very much in line with current understandings of validity theory. Lastly, validation is viewed as a *sequential* process where guidance is given on when different types of validity evidence should be gathered in relation to different stages of the test cycle. Weir (2005) also provides suggestions for various methods of collecting validity evidence for each component in the framework which altogether provide a systematic way of collecting validity evidence. The potential for operationalising the framework is one of its major strengths and is further evidenced in the various projects in which it has been successfully applied. (see Weir and Khalifa, 2009; Taylor, 2011; O’Sullivan and Weir, 2011 for examples of projects which have applied the model).

2.4.4 Application of the Weir (2005) Framework to the Current Study

This section provides a brief description of the elements in Weir’s (2005) framework before explaining how the foci of the current study, i.e. topic and background knowledge of topic fit into the framework.

The *test taker* box in the framework is directly linked to the theory-based validity box and together, they represent the individual candidate in a test (O’Sullivan and Weir, 2011). Drawing on the work of O’Sullivan (2000a:82-82) test-taker characteristics are divided by Weir (2005) into three groups (physical, psychological and experiential) and it is hypothesised that different features within each group are likely to be in constant

interaction with one another. *Physical/physiological* characteristics refer to ‘fixed’ biological features such as gender and age and other characteristics such as short-term ailments and longer-term disabilities; *psychological characteristics* are those relating to aspects internal to the test taker (O’Sullivan and Green, 2011) such as personality, affective schemata (Bachman and Palmer’s 1996), and motivation; *experiential characteristics*, on the other hand, refer to those aspects external to the test taker such as education, language learning and exam preparation experience, etc. In O’Sullivan and Green (2011:40) ‘general world knowledge/ topic knowledge’ is also subsumed under experiential characteristics as an important feature likely to influence test performance and one which should be taken into account when designing tasks. Individuals from different age or cultural backgrounds can therefore be expected ‘to have different kinds of knowledge about the world’.

These test-taker characteristics can influence the way in which a task is performed. It is important to demonstrate that characteristics of the test takers that are irrelevant to construct under examination do not have a significant impact on performance. Also relevant is a consideration of ‘differential item functioning’ (DIF) or test bias which refers to the ‘presence of some characteristic of an item that results in differential performance for individuals of the same ability but from different ethnic, sex, cultural or religious groups’ (Hambleton and Rodgers, 1995:1). In Weir (2005), DIF was considered under consequential validity. However, in more recent work Taylor (2011:30) argues for DIF to be placed under test-taker characteristics, as ‘the evidence collected on the test taker should be used to check that no unfair bias has occurred for individuals as a result of decisions taken earlier with regard to contextual features of the test’. In line with this

argument, if ‘topic’ is regarded as a contextual feature of a test and ‘background knowledge of topic’ is regarded as a test-taker feature, then it is necessary for upholding topic validity to demonstrate that no unfair bias has occurred against individuals as a result of the topic of tasks. In examining bias, Bachman (1990) cautions that observed differences in performances may not always necessary reflect bias but differences in language ability, which also explains why many of the studies reviewed in the literature attempt to control for general language proficiency in evaluating group differences.

Under the test taker element of the Weir (2005) model, on the basis of the reviewed literature, and in line with the requirement for test fairness and validity for test takers from different backgrounds and with different characteristics, the following research questions are formulated, the answer to which can provide test taker-related validity evidence:

- *Will the differences in test takers’ levels of background knowledge of topics have an impact on performance? Are (any) differences large enough to have practical significance in terms of test performance?*
- *Does background knowledge of topics differentially affect performances of test takers from different proficiency levels?*
- *What are the test takers’ perceptions of the influence of topic and background knowledge of topic on performance?*

The *context validity* box in the framework is used as a ‘superordinate category’ related to tasks and performance conditions under which these speaking tasks are performed (Taylor, 2011). Within the test constraints, Weir (2005: 56) suggests that performance conditions should be as closely similar to authentic language use as possible. The elements in the context validity box include characteristics of the task setting, the demands of the task both in terms of linguistic input and output, characteristics of the interlocutor (and

number of interlocutors) as well as the administrative setting of the test event. Topic familiarity/content knowledge falls under the linguistic demands of the task. Weir (2005) also attributes a significant role to the topic of a task and how that ‘the content knowledge required for completing a particular task will affect the way it is dealt with’. For general proficiency tests, it is recommended for ‘neutral topics’ to be selected so as not to favour any particular group of people (Galaczi and French, 2011:149). Weir (2005:76) also draws attention to a concern that different topics may elicit ‘responses that are *measurably* different’. In direct response to this concern, the following questions are formulated under the context validity component of Weir’s (2005) framework in relation to task topics in the IELTS Speaking Test:

- *When task type is held constant, to what extent are the different topics used in parallel versions of a task similar in terms of difficulty? Are (any) differences in topic difficulty measures large enough to have practical significance in terms of test performance?*
- *When task type is held constant, to what extent are the observed functions elicited by different topics similar?*

The *cognitive validity* component (previously referred to as theory-based validity) relates to the mental processes activated by the tasks in the context validity box. The elements in the box should reflect an established theory of the mental processes underlying the construct of interest. For a speaking test, therefore, the cognitive validity evidence should be able to demonstrate that speaking tasks ‘activate cognitive processes in a test taker’s mind similar to those employed in the real-life speaking tasks and situations the test aims to generalize to’ (Weir, Vidakovic and Galaczi, in press). Weir’s (2005) cognitive validity component for speaking is comprised of executive processes, executive resources and

monitoring features which reflect Levelt's (1989) widely recognised model of L1 speech. More recently, the Levelt (1989) model was re-visited by Field (2011:75) who looked at how different task conditions could influence different stages of cognitive processing and identified the ways in which these differences could be potentially captured. A full discussion of these cognitive processes is beyond the scope of this section. However, attention is drawn to the 'conceptualisation' stage of speech processing (from Levelt's model) which is concerned with the generation of ideas, amongst others and is thus the most relevant aspect of the cognitive validity component to the current study. The generation of ideas is viewed as a cognitively demanding process, which can increase as a result of the complexity of ideas. Complexity, in this case, can be seen on a continuum from concrete to abstract, familiar to unfamiliar and personal to non-personal ideas/information along which tasks can vary in terms of the cognitive demand they pose on the test taker (Field, 2011; Weir, Vidakovic and Galaczi, in press).

It was explained in Chapter 1 that the IELTS Speaking Test is comprised of three parts which are designed to increase in difficulty in line with assumptions of decreasing familiarity of topics and provision of ideas where Part 1 is the easiest due to the use of familiar topics with Part 3 topics as the most difficult due to the abstract nature of the topics with Part 2 topics falling in between due to the provision of background knowledge/ideas in the form of prompts (see section 1.6.2). Task topics are therefore designed to pose increasing levels of cognitive demand on the test taker as a function of familiarity and abstractedness. The observed sequence of topic difficulties matched against the intended sequence of tasks may provide evidence of whether different topics

used in a test demonstrate cognitive validity. The following research question is formulated:

- *To what extent does the observed progression of topic difficulty measures match the intended progression of topic difficulty from easy to difficult?*

Scoring validity is a superordinate term used by Weir (2005:22) to refer to all aspects of reliability and include considerations of rating criteria and rating scales and the application of the rating scales by human raters (or computers in the case of automated marking of speaking test). It is important to ascertain that there is an acceptable degree of agreement between raters and that procedures are in place to minimise measurement error, for example, through rater training and/or the use of statistical procedures such as MFRM. Taylor (2011:29) interprets scoring validity as ‘the extent to which test scores are arrived at through the application of appropriate criteria and rating scales by human judges, as well as the extent to which they exhibit agreement, are as free as possible from measurement error, stable over time, appropriate in terms of their content sampling and engender confidence as reliable decision-making indicators’.

For the current study, it is necessary to bring forward evidence of scoring validity in order to demonstrate that interpretations regarding the effects of topics and background knowledge of topics are not influenced by systematic rater tendencies. These are general considerations for any rater-mediated study. However, it was suggested in the literature that raters may interact with other facets of the assessment context. For example, raters may favour candidates who exhibit higher background knowledge or may exhibit bias towards particular topics. Relevant to the current study, the following questions are posed in relation to scoring validity:

- *Are there any significant interactions between raters and a) topics, b) test takers' background knowledge of topics and c) test takers' language proficiency levels?*

Consequential validity was discussed to a large extent earlier (see section 2.4.1). It refers to a) the impact of tests on institutions and society b) washback in the classroom or workplace and c) test bias. However, as previously noted, Taylor (2011:30) argued for a consideration of test bias to be subsumed under test-taker characteristics. For this reason and due to the scope of the study, other aspects of consequential validity are not addressed in the current study and test bias is discussed in relation to the test taker element of the model.

Lastly, under *criterion-related validity*, Weir (2005) invites us to consider evidence, external to the test, which can lend further support for the meaningfulness of scores or ‘the extent to which test scores reflect a suitable external criterion of performance or demonstration of the same abilities as are included in the test (O’Sullivan and Weir, 2011:23-24). Three forms of evidence have been identified by Weir (2005) and Khalifa and Weir (2009) which can lend to support to a criterion-related validity. Firstly, evidence of a relationship between test scores and an appropriate external criterion with established properties (Anastasi, 1988). Secondly, evidence collected from linking a given test to an external standard through systematic procedures. A ubiquitous example is the aligning of language tests to the Common European Framework of Reference (see Bachman, 2011 for some of the issues raised regarding linking and aligning interpretations across frameworks). Thirdly, evidence which demonstrates ‘the qualitative and quantitative equivalence of different forms of the same test’ (Khalifa and Salamoura, 2011:259). In relating the latter to the current study, if different forms of a speaking test consist of

different combination of topics (within task types), then a piece of evidence for criterion-related validity can be derived from evidence of topic comparability. The following question is therefore posed under the criterion-related validity of the framework in relation to topics:

- *To what extent are parallel forms of a language proficiency interview (consisting of different topics) comparable in terms of difficulty? Are (any) differences in form difficulty large enough to have practical significance in terms of test performance?*

It should be noted that under different validity components of the model, numerous combination of questions could be posed. This is a strength of the framework, as it allows for the examination of any number of hypothetical interactions between different features of the framework. Similarly, different kinds of evidence can be gathered to answer each question. The types of evidence gathered will reflect the nature of the questions asked, and the scope of a validation project will depend on limitations of single projects. For the current project, a construct-irrelevant validation approach was considered a priority and several research questions were formulated accordingly (see section 1.8). The next chapter details the ways in which validity evidence is gathered to answer the questions while taking into account the project's practical limitations.

Chapter 3 Methodology

This chapter presents and justifies the research design, methodologies and procedures undertaken for addressing the following overarching research question:

How is the topic validity of a test of second language speaking performance influenced by the random assignment of topics to test takers who bring different levels of background knowledge to the topics?

Answering this question necessitates the collection of different types of validity evidence.

The question was therefore operationalised into the following research questions.

- i. To what extent are the topics (speaking tasks) used in parallel versions of a language proficiency interview similar in terms of difficulty?*
- ii. To what extent does the observed progression of topic difficulty measures match the intended progression of topic difficulty from easy to difficult?*
- iii. To what extent are parallel forms of a language proficiency interview (consisting of different topics) comparable in terms of difficulty? Are (any) differences in form difficulty large enough to have practical significance in terms of test performance?*
- iv. When task type is held constant, to what extent are the different topics used in parallel versions of a task similar in terms of difficulty? Are (any) differences in topic difficulty measures large enough to have practical significance in terms of test performance?*
- v. When task type is held constant, to what extent are the observed functions elicited by different topics similar?*
- vi. Will the differences in test takers' levels of background knowledge of topics have an impact on performance? Are (any) differences large enough to have practical significance in terms of test performance?*

vii. *Does background knowledge of topics have an impact on topic difficulty measures?*

viii. *Does background knowledge of topics differentially affect performances of test takers from different proficiency levels?*

ix. *Are there any significant interactions between raters and a) topics, b) test takers' background knowledge of topics and c) test takers' language proficiency levels?*

x. *What are the test takers' perceptions regarding the influence of topic and background knowledge of topic on performance?*

xi. *What themes and patterns emerge from a qualitative inquiry into a) rater perspectives and b) the content of test taker speaking performances, in relation to topic validity aspects of the speaking test under examination?*

3.1 Approach to Research

This study employs a mixed-methods approach to research where both quantitative and qualitative methods and strategies are adopted as appropriate. The strategies of inquiry fit in with Creswell and Plano-Clark's (2007:71) 'embedded design' in which, similar to a 'concurrent triangulation design', both quantitative and qualitative strands of research are conducted and 'merged' during the interpretation of the study's findings. However, in an *embedded design*, the two strands of inquiry may not hold equal weighting and one form of data is often embedded within a larger data collection procedure. An embedded design is considered appropriate when 'the researcher has different questions that require different types of data in order to enhance the application of a quantitative or qualitative design to address the primary purpose of the study' (Creswell and Plano-Clark, 2007:91).

For the current study, the majority of subsidiary research questions are addressed using quantitative techniques, involving the analysis of score data and questionnaire closed-

ended response data. However, the study's final subsidiary research question involves a qualitative analysis of the content of test-taker performances and raters' interview transcripts. The data analyses of the two strands are inquiry are carried out independently but are 'mixed' and synthesised in the interpretation phase and in addressing the study's overarching research question. In investigating the effects of topic and background knowledge of topic on performance, the validity argument lies predominantly on quantitative data. However, the qualitative information can enrich research findings and provide a deeper understanding of the research problem (Creswell, 2003; Tashakkori and Teddlie, 2002).

3.2 Research Setting

There has been a recent move towards the 'localisation' of international tests where it is argued that international tests should also demonstrate local validity (O'Sullivan, 2011:6). For the current study, the local context of Iran was selected as the research setting for the following reasons:

- The researcher's familiarity with the Iranian context
- Iranian candidates are amongst the top 20 most frequent test-taking nationalities (IELTS, 2007)
- Despite the frequency of IELTS test administrations, no studies (to the best of my knowledge) have been carried out in the Iranian context of IELTS. Such a study is in line with the IELTS research programme agenda (2009-2010) which highlighted the need for studies to be carried out in local under-researched contexts.

3.3 Data Collection (Phase I)

The study's data collection involves two main phases: data collection from test-taker participants in phase I, and the scoring of test-taker performances by raters in phase II. Each phase is described separately, as they entail different participants, instruments and procedures.

3.3.1 Participants (Test Takers)

A total of 82 adult non-native speakers of English (L1 Farsi), aged between 18 and 40, participated in this study. All test takers were enrolled in English language courses and were preparing to take the IELTS exam. They were familiar with the general format of the Speaking Test as confirmed by their teachers. The gender of the participants was evenly distributed.

3.3.2 Instruments

3.3.2.1 Speaking Tasks/Speaking Test Forms

3.3.2.1.1 *Number of Speaking Tasks/Parallel Forms*

Prior to selecting the study's speaking tasks, it was important to establish the number of tasks/topics that could be feasibly included in the study; on the one hand there was a desire to include as many topics as possible and on the other, there was a limit on the number of topics that each individual participant could respond to without inducing fatigue or boredom.

A decision was made to opt for a parallel forms reliability design where each participant responds to two parallel/alternative versions of the IELTS Speaking Test. To remind the reader, each Speaking Test is comprised of three task types (A, B and C) with two Task A

topics, one Task B topic and two Task C topics, that is, a total of five topics per Speaking Test (see section 1.6.2). By adopting a parallel forms design, each participant responds to a total of 10 topics. However, instead of having all participants respond to the same two parallel Speaking Tests i.e. a *complete* or *fully crossed design* (Eckes, 2009), a more practical solution was found in an *incomplete design* (Eckes, 2009; Weir and Wu, 2006). An incomplete design necessitates the *linking* of the tests through common tasks but simultaneously allows for an increase in the number of topics that can be included. It should be noted that this design-related decision was made in light of the measurement model selected for the study, that is, MFRM, as the model is robust against missing data as long as there is enough connectedness or linking between the elements of different facets. Eckes (2009:39) defines a connected data set as ‘one in which a network of links exists through which every element that is involved in producing an observation is directly or indirectly connected to every other element of the same assessment context’.

In this study, a decision was made to construct four alternate versions of the IELTS Speaking Test with two common tasks creating the necessary *common link* between the tests, allowing for coverage of 18 different topics. Table 3-1 illustrates the incomplete – connected data collection design of the study. Topics A.1 and A.2 (in bold) are the common items.

Table 3-1 Illustration of the Incomplete –Connected Data Collection Design in the study

TOPICS	TASK TYPES	EXAMINEE (GROUP 1)	EXAMINEE (GROUP 2)
A.1	A	X	X
A.2	A	X	X
A.3	A	x	
A.4	A	x	
A.5	A		x
A.6	A		x
B.1	B	x	
B.2	B	x	
B.3	B		x
B.4	B		x
C.1	C	x	
C.2	C	x	
C.3	C	x	
C.4	C	x	
C.5	C		x
C.6	C		x
C.7	C		x
C.8	C		x
Test Form W: A.1, A.2 , B.1, C.1, C.2 Test Form X: A.3, A.4, B.2, C.3, C.4 Test Form Y: A.1, A.2 , B.3, C.5, C.6 Test Form Z: A.5, A.6, B.4, C.7, C.8			

3.3.2.1.2 *Speaking Task Selection*

In selecting speaking tasks, two main steps were taken. First, a review of previously published IELTS examination papers (retired material) was carried out. As Weir (2005) and Weir and Wu (2006) emphasise, it is integral to upholding the validity of a test to demonstrate that tasks and test forms used across administrations and years have similar difficulty levels, particularly in the case of large-scale high stakes tests IELTS and TOEFL. The retired IELTS materials consist of authentic tasks which have been previously used in live operational settings and therefore serve the purposes of the study. Following a preliminary review of the available material, a total of eight Task A topics, five Task B topics and 10 Task C topics (thematically linked to task B topics) were

selected. Secondly, a more in-depth analysis of the tasks was carried out in order to establish task equivalence.

3.3.2.1.3 Establishing Task Equivalence

One of the aims of this study is to examine the extent to which topic variance has an influence on scores. In line with Bachman (2002) and Weir, O'Sullivan and Horai (2006), it is important to ascertain that with the exception of task topic, other task-related variables are controlled for as far as possible so that (any) differences in scores could be predominantly attributed to differences in task topics and test takers' background knowledge of the topics.

The literature review (see section 2.1.1) identified the task processing conditions of code complexity, cognitive complexity and communicative stress (Skehan and Foster, 1997) as factors which can affect second language performance. Weir, O'Sullivan and Horai (2006:5) break down these factors to specific features of 'planning time, planning condition, audience, type and amount of input, response time and topic familiarity'. With the exception of **topic familiarity**, which constitutes the focus of the study and the condition which varies across tasks, the remaining conditions have to be addressed and controlled for.

Planning time, **planning condition**, and **response time** can be controlled through test administration procedures and by adhering to a strict interview structure. For example, providing a specific frame for each part of the test controls for response time while for Task Type B (Part 2), the provision of a one-minute preparation time and the written prompts controls for planning time and planning conditions respectively. In this study, the examiner/interlocutor role was fulfilled by the researcher in order to control for the

potential effects of **audience** and **interlocutor-related features** on performance. Adherence to the IELTS Speaking Test format also implies that **type of input** in different parts of the test is held constant for all participants.

Amount of input was both quantitatively and qualitatively examined in order to ensure task equivalence in relation to level of input.

The quantitative analysis of the tasks involved the calculation of descriptive statistics in terms of average number of questions per task, words per sentence, and characters per word for all tasks (see results in Appendix D). For Task B tasks (written prompts), the Flesch-Reading Ease (Flesch, 1948) and Flesch-Kincaid grade level (Kincaid, 1975) were also calculated and taken into account in the selection process following Weir, O'Sullivan and Horai (2006).

The qualitative examination of the tasks was carried out by eliciting the perceptions of a select panel of colleagues (from applied linguistics/second language acquisition academic backgrounds) regarding the extent of task equivalence. The panel was asked to apply a task equivalence checklist (adapted from Weir, O'Sullivan and Horai, 2006 and Weir and Wu, 2006) to different tasks, rate the topics within each task type on a number of different criteria (e.g. lexis, grammar, functions and topic of the tasks) and to provide extended comments on tasks (see Appendix C for a copy of the checklist).

The final task selection decisions were made on the basis of a number of different types of evidence i.e. the panel's ratings of tasks, the qualitative analysis of task-related comments and a consideration of task input statistics. Tasks which were flagged for containing relatively unfamiliar grammar or lexis, were removed. On the other hand, tasks which were considered to be relatively more unfamiliar or abstract in terms of topic and content

were retained. In fact, the variance in the panel's perceptions of topic familiarity across tasks was a desirable outcome, as it suggested that a mix of familiar and unfamiliar topics was selected for the final study. This analytic exercise allowed for the selection of task topics which exhibited equivalence in terms of features to an acceptably high degree. The final set of speaking tasks employed in the study are presented in Figure 3-1, Figure 3-2, and Figure 3-3 for Task Types A, B and C respectively.

These tasks were then divided into four alternate Speaking Test forms (W, X, Y and Z) according the IELTS Speaking Test format. These forms, the sequence of task administration, the response times, and the speaking test scripts can be accessed in the appendices (Appendix E to Appendix H).

<u>Topic A.1 Family</u> Let's talk about your family. • Do you have a large family or a small family? • Can you tell me something about them? • How much time do you manage to spend with members of your family? • What sorts of things do you like to do together? (UCLES, 2002:29)	<u>Topic A.2 :Leisure Time</u> Now let's talk about your leisure time. • Do you have any hobbies or interests? [What are they?] • What is there to do in your free time your city? • How do you usually spend your holidays? • Is there anywhere you would particularly like to visit? [Why?] (UCLES, 2005:80)
<u>Topic A.3: Festivals</u> Let's talk about festivals. • Tell me about the most important festival in your country. • What special food and activities are there in this festival? • What do you enjoy most about it? • Do you think festivals are important for a country? (UCLES: 2002:53)	<u>Topic A.4: Colour</u> Now let's move on to talk about colour. • What is your favourite colour? [Why?] • Do you like the same colours now as you did when you were younger? [Why/Why not?] • What can you learn about a person from the colours they like? • Do any colours have a special meaning in your culture? (UCLES, 2006:54)
<u>Topic A.5 : Keeping in contact with people</u> Let's talk about keeping in contact with people. • How do you usually contact your friends? [Why?] • Do you prefer to contact different people in different ways? [Why?] • Do you find it easy to keep in contact with friends and family? [Why/Why not?] • In your country, did people in the past keep in contact in the same way as they do today?[Why/Why not?] (UCLES, 2009: 32)	<u>Topic A.6: Dancing</u> Now let's move on to talk about dancing. • Do you enjoy dancing? [Why/ Why not?] • Has anyone ever taught you to dance? [Why/Why not?] • Tell me about any traditional dancing in your country. • Do you think that traditional dancing will be popular in the future? [Why/Why not?] (UCLES, 2008: 32)

Figure 3-1 Task Type A Topics Selected for the Main Study

<u>Topic B.1: Describe one of your friends</u> Describe one of your friends. You should say: - How you met - How long you have known each other - How you spend time together And explain why you like this person. (UCLES, 2006:77)	<u>Topic B.2: Describe a river, lake or sea</u> Describe a river, lake or sea which you like. You should say: - What the river, lake or sea is called - Where it is - What the land near it is like And explain why you like this river, lake or sea. (UCLES, 2005:80)
<u>Topic B.3: Describe an important choice you had to make</u> Describe an important choice you had to make in your life. You should say: - When you had to make this choice - What you had to choose between - Whether you made a good choice And explain how you felt when you were making this choice. (UCLES, 2008: 100)	<u>Topic B.4 : Describe someone in your family</u> Describe someone in your family who you like. You should say: - How this person is related to you - What this person looks like - What kind of a person he/she is And explain why you like this person. (UCLES, 2008: 32)

Figure 3-2 Task Type B Topics Selected for the Main Study

<u>Topic C.1: Qualities of friends</u> • What do you think are the most important qualities for friends to have? • Which are more important to people, their family or their friends? [Why?] • What do you think causes friendships to break up? (UCLES, 2006:77)	<u>Topic C.2: Other relationships</u> • What other types of relationship, other than friends or family, are important in people's lives today? • Have relationships with neighbours where you live changed in recent years? How? • How important do you think it is for a person to spend some time alone? [Why/Why not?] (UCLES, 2006:77)
<u>Topic C.3: Water-based leisure activities</u> • What do people enjoy doing when they visit rivers, lakes or the sea? • What benefits do you think people get from the activities they enjoy in the water? • What are the different advantages and disadvantages of going to the sea or to a swimming pool to enjoy yourself? (UCLES, 2005:80)	<u>Topic C.4: The economic importance of rivers, lakes & the sea</u> • How does water transport, like boats and ships, compare with other kinds of transport? • How important is it for a town or city to be located near a river or the sea? Why? • Have there been any changes in the number of jobs available in fishing and water transport industries, do you think? [Why] (UCLES, 2005:80)
<u>Topic C.5: Important Choices</u> • What are the typical choices people make at different stages of their lives? • Should important choices be made by parents rather than by young adults? • Why do some people like to discuss choices with other people? (UCLES, 2008: 100)	<u>Topic C.6: Choices in everyday life</u> • What kind of choices do people have to make in their everyday life? • Why do some people choose to do the same things every day? Are there any disadvantages in this? • Do you think that people today have more choices to make today than in the past? (UCLES, 2008: 100)
<u>Topic C.7: Family Similarities</u> • In what ways can people in a family be similar? • Do you think that daughters are always more similar to mothers than to male relatives? What about sons and fathers? • In terms of personality, are people more influenced by their family or their friends? In what ways? (UCLES, 2008: 32)	<u>Topic C.8: Genetic Research</u> • Where can people in your country get info about genetic research? • How do people in your country feel about genetic research? • Should this research be funded by governments or private companies? Why? (UCLES, 2008: 32)

Figure 3-3 Task Type C Topics Selected for the Main Study

3.3.2.2 Background Knowledge Questionnaire

3.3.2.2.1 *Assessing Background Knowledge of Topics*

The literature review outlined the various methods with which the construct of background knowledge or topic familiarity has been operationalised in research in the field (see 2.1.4). For example, some authors argued that level of familiarity should be ‘clear-cut’ (Bei, 2010: 64) and that there is a need to ‘impose hard criteria’ (*Ibid.*). In the literature such hard criteria mostly included extreme examples such as matching/mismatching of individuals to different culture-laden topics or to very specific academic backgrounds (Bei, 2010). In many examples, background knowledge was either assumed, or inferred on the basis of academic field of study, cultural background etc. Other approaches included establishing familiarity using binary Yes/No questions or using topic knowledge tests. Jennings et al. (1999) referred to the complexities involved in determining background knowledge given the influence of multiple factors such as interest in topic, affective schemata and ideas about a topic and therefore opted for a choice mechanism in capturing these complexities.

The approach taken in this study follows from Jennings et al. (1999) in that the interaction between a test takers’ background knowledge and topic of a task is viewed as a complex phenomenon which cannot be assumed or pre-determined. On the other hand, simply relying on a ‘choice’ mechanism does not allow for a thorough examination of the role of background knowledge. For this reason, a decision was made to construct a background knowledge questionnaire with the aim of capturing *relative* degree of background knowledge of topics while taking the possible influence of other topic-related factors into account, for example, by including questions on topic-related interest and perceptions of

topic difficulty. It was also posited that participants, having responded to a range of different topics (through the parallel forms reliability design), would be able to gauge their relative degree of background knowledge for the given topics. Post-speaking test questionnaires were therefore constructed to elicit test takers' self-reports of background knowledge of given topics.

3.3.2.2.2 *Background Knowledge Questionnaire*

In this section, the different parts of the designed questionnaire and the rationale behind their inclusion are described. The questionnaire is divided into three main sections:

Section I

This section consists of eight questions (reproduced in Table 3-2), repeated for each topic. These questions were formulated on the basis of the literature and following a review of available instruments (e.g. Weir, 2005; Weir, O'Sullivan and Horai, 2006). Responses are elicited on a five-point Likert scale (1=Strongly Disagree to 5=Strongly Agree; 3=Undecided). A space is provided for open-ended comments.

Table 3-2 Topic-related Questions

Topic-related Questions
1. This topic is familiar to me.
2. The questions about this topic were easy to respond to.
3. I know <i>A LOT</i> about this topic, i.e., I have <i>MORE THAN ENOUGH IDEAS</i> to talk about this topic.
4. It was easy for me to produce enough ideas for this topic from memory.
5. If I were to talk about this topic in Farsi, I would have <i>MORE IDEAS</i> to talk about.
6. I had appropriate words to express my ideas about this topic easily.
7. I thought this was an interesting topic.
8. I performed very well on this task.

Some questions are designed to elicit participants' self-reports of degree of topic familiarity and ideas about a topic (more directly associated with background knowledge) while other questions elicit other topic-related aspects such as interest in topic (in line with Carrell and Wise, 1998, and Jennings et al., 1999) and perceived difficulty of topic. Question 5 attempts to tease language ability and background knowledge apart. Questions 2, 3, 4 and 6 were adapted from Weir's (2005:237-239) cognitive processing questionnaire. For example it was hypothesised that a higher degree of familiarity/background knowledge would correspond with an easier production of ideas from memory (Question 4).

Given that each participant responded to 10 different topics across the two speaking test forms, this section of the questionnaire consisted of 80 questions (10 topics x 8 questions).

Section II

This small section of the questionnaire focuses on test-taker perceptions of the role of topic and topic familiarity on their performance and the importance they place on these factors. The four questions in this section are reproduced below:

For each of the items below, circle the number that reflects **your viewpoint** on a five-point scale where:

1= Strongly disagree, 2= Disagree, 3= Undecided, 4= Agree, 5=Strongly Agree

1. I think that the choice of topics might affect my final score.	1	2	3	4	5
2. I think that having more ideas about a topic might affect my final score.					
3. I think that there is an element of 'luck' involved in the choice of topics.					
4. I think that the choice of topic is not important if my English is good enough.					
-

While questions 1 and 2 focus attention on the effects of topic and background knowledge of topic on scores, question 4 aims to examine whether the inclusion of ‘language ability’ has an effect on the pattern of responses in relation to topic effects on performance. These questions are intended to tap into test-taker perceptions of test fairness, as none of these variables i.e. choice of topic, test takers’ background knowledge of topics or the ‘luck’ of the draw in terms of topic should be perceived as having a significant effect on speaking performance.

Section III

This final section simply elicits some general information from participants (e.g. gender, education, self-reports of second language proficiency, familiarity with the test format). This section is placed last, as it poses less cognitive demand on the participants, compared to the other parts, since it only requires factual information.

3.3.2.3 C-tests

3.3.2.3.1 Assessing General Language Proficiency

The literature review suggested general language proficiency to be one of the variables that can shape the way in which a test taker’s background knowledge of a topic interacts with a task topic. In order to examine the potential interaction between background knowledge of topic and language proficiency, a measure of general language proficiency was required. A decision was made to opt for C-tests for this purpose, following an in-depth review of the literature on C-tests which strongly supported their construct validity¹⁹. In the words of Eckes and Grotjahn (2006: 300) there is ‘converging evidence

¹⁹ Due to space limitations, the extended literature review on C-tests is not provided in this section and only a brief overview is presented.

for the suitability of C-tests as measures of general language proficiency'. It is also the relative ease of development, administration and scoring of C-tests that made it an appealing option as a research instrument and a control measure for general language proficiency. In the next sections, a brief introduction to C-tests is presented, followed by the steps taken in the development, piloting and modification of the study's C-tests.

3.3.2.3.2 *What is a C-test?*

A C-Test is defined as an 'integrative' written test of general language proficiency (Raatz and Klein-Braley, 2002) and belongs to a branch of language tests that are built on the principle of reduced redundancy (Spolsky, 1981). The assessment of language ability in such tests is based on the extent to which test takers draw on their knowledge of the target language to restore linguistic messages that have been distorted in one way or another and usually through a systematic introduction of an element of 'noise' or 'interference' to the original message (Babaii and Ansary, 2001:210).

3.3.2.3.3 *The Construct Validity of C-tests*

The views on what C-tests actually measure and their construct validity are divided. On the one hand are those who view C-tests as having limited functionality in assessing micro-level processing only (Cohen et al., 1984; Kamimoto, 1992). This is explained by the hypothesis that half-deleted words can elicit lexical items without relying on contextual clues and macro level processing (Cohen et al., 1984) and that is the knowledge of grammar and vocabulary that is predominantly targeted by these types of test (Kamimoto, 1992). On the other hand there is empirical evidence which is in contrast with the above findings and explains performance on the C-tests as a function of both top-down and bottom-up processing (Feldmann and Stemmer, 1987) and as a highly integrative test

of language (Babaii and Ansary, 2001). In their comprehensive review of the literature on C-tests, Eckes and Grotjahn (2006) survey studies which had adopted various methods in determining the construct validity of C-tests ranging from quantitative methods of factor analyses, correlations to more qualitative measures such as verbal protocol. In their own study, Eckes and Grotjahn (2006) employed two strong statistical analytical approaches of Rasch analysis and confirmatory factor analyses in conjunction and used a standardised large scale test of German as an independent criterion of proficiency in order to investigate the construct validity of C-tests. Their main hypothesis was that the trait underlying performance on the C-tests is the same as that underlying performance on TestDaF and its subsections. Findings from their study supported the hypothesis:

Rasch measurement modelling and confirmatory factor analysis provided clear evidence that the C-test in question was a highly reliable, unidimensional instrument, which measured the same general dimension as the four TestDaF sections: reading, listening, writing and speaking. Moreover, the authors showed that language proficiency was divisible into more specific constructs and that examinee proficiency level differentially influenced C-test performance.

Eckes and Grotjahn (2006: 290)

On the basis of the literature, evidence of high correlations between C-test scores and the productive skill of speaking (Singleton and Singleton, 2002) and the observation that ‘the value of C-testing as a measure of global proficiency in second language has been demonstrated too many times to be open to dispute’ (Hastings, 2002 cited in Eckes and Grotjahn, 2006: 292), C-tests were selected as a suitable measure of general language proficiency.

3.3.2.3.4 *C-test Design and Piloting*

In designing the study's C-test, several resources (Klein-Braley and Raatz, 1984; Klein-Braley, 1997) were first consulted for guidelines and instructions regarding C-test construction. The following steps were subsequently taken²⁰:

- Text selection (a total of 15 texts were originally selected)
- Establishing text difficulty using the Lexile-measure²¹
- Text conversion to C-test format
- C-test piloting (with all texts) with native speakers of English
- Modification/elimination of texts following pilot study results
- Piloting with non-native speakers of English
- Selecting texts for the final C-test
- Developing an answer key on the basis of the pilot study responses

The three versions of the C-test can be accessed in Appendix I. A common-item linking strategy was also adopted for the three versions of the C-tests in order to include a larger selection of texts and to increase reliability.

3.3.3 Procedures

3.3.3.1 Participant Selection Process

A total of 12 language schools in Tehran, which offered IELTS test preparation courses, were contacted for collaboration in the research project via e-mail and telephone. The e-mail contained general information regarding the study e.g. purpose of the project, what participation in the project would entail and data collection schedule. A sample e-mail is

²⁰ The original section on C-test design, development and piloting could not be included due to space limitations. In this section, a summary of the steps taken is presented.

²¹ The Lexile measure (or L-measure) is a Rasch calibrated measure, which gives objective information about the difficulty of a text or a person's reading ability. It is calculated on the basis of word frequency and sentence length, as the two strong predictors of text difficulty (MetaMetrics, 2011)

provided in Appendix J. The Oxford's Central University Research Ethics Committee (CUREC) approval document was also attached to the e-mail.

Of the twelve institutes, only five responded positively and showed an interest in collaboration. This number was further reduced to two institutes as the other three required the proposal to go through a time consuming bureaucratic procedure which would conflict with the scheduled plans. In addition to the language schools, two teachers (former colleagues of the researcher) offering semi-private courses for IELTS exam preparations were also contacted for access to their students.

Once approval for data collection was received from the language schools and EFL teachers, the students enrolled in exam preparation courses were briefed (during class hours) about the study although the exact focus of the study was not disclosed in order to allow for a spontaneous speaking test performance. The students were told that would have the opportunity to practice two mock speaking tests (with an IELTS format) with a former IELTS examiner and that they would receive feedback regarding their performance. They were also informed that the speaking tests would be recorded and that all personal information would be anonymised. Participant Information Sheets were distributed amongst classes (see Appendix K). The students were very eager to participate in the study, as it provided them with extra speaking practice for their future exams and they gave their consent for participation (see Appendix L for a sample consent form). Participation was strictly on a voluntary basis and no honorariums were provided.

3.3.3.2 Data Collection Procedure

Data collection was planned to take place during normal class hours at the language schools. The sequence of instrument administration was as follows:

- C-test administration
- Speaking test administration
- Questionnaire administration

3.3.3.2.1 *C-test Administration*

The results of the C-test pilot study suggested a completion time of 30-35 minutes. Participants were invited to arrive at the language schools/classes 40 minutes prior to their scheduled class time, and were presented with short instructions on how to complete the C-tests. Additionally, they were told that the test might prove to be difficult and that they should not feel discouraged if they could not fill in the blanks. They were also advised to revise the texts – time permitting. The three forms of the C-test were randomly distributed and administered to the participants present at each class, and were then collected at the end of the allotted time.

3.3.3.2.2 *Speaking Test Administration*

The administration of the speaking tests had to take place parallel to the participants' English class schedules, which ran between 2 and 3 hours. In each testing session, a maximum of 6 persons participated, as speaking test administration took approximately 30 minutes per individual. It was agreed in advance that participants would be called out one by one to a quiet room, where the speaking tests would be administered by the researcher.

Two forms of the speaking tests (either W and X or Y and Z, see Appendix E to Appendix H) were administered to participants by random rotation. The order of administration was

also alternated to control for an order effect. The administration of the two forms of the speaking tests took place in succession. It would have been preferable to administer the tests on two different occasions so that the interpretation of scores from the test administrations would not be limited across forms but also across occasions (Anastasi, 1988). Weir and Wu (2006) suggest the following:

The administration of parallel (alternate) forms of a test in independent sessions provides us with alternate-form coefficients. The two tests are normally given to an appropriate single group of learners with a short period between the two administrations. A 'short' period would be not so short that the learners would be exhausted or bored by the process, and not so long that significant learning might take place between the two administrations. This will measure both the temporal stability of the test and the consistency of response to the two samples of test tasks.

Weir and Wu (2006:169)

In Khabbazzbashi (2008), which served as a pilot for the current study, the data collection process aimed for the administration of forms on two different occasions. This proved to be very difficult due to the no-show rates of participants on both occasions (on account of traffic, illness, personal schedules, etc.). A decision was therefore made to administer tests in succession. The participants were subsequently asked to comment on the two factors of exhaustion and boredom, pointed out by Weir and Wu (2006), as possible (negative) outcomes of the modified procedure. No complaints were made about fatigue, and participants also commented that the different topics of the two tests kept them interested.

Following the pilot study and in order to increase the chances of collecting a complete data set and to ensure minimum disturbance to the classroom schedules of participants, a practical decision was made to administer the two speaking tests in succession in the current study. This decision also allowed the participants to draw immediately on their test-taking experience of all the task topics when filling in the post-test questionnaires.

Prior to the administration of speaking tests, the researcher provided the participants with a brief description of the format of the speaking tests and the duration of each test. All speaking tests were digitally recorded. At the start of recording, each participant was identified with a reference code. All recorded data files were later transferred to a computer.

3.3.3.2.3 Questionnaire Administration

Following each testing session, the participants were asked to stay an additional thirty minutes after the end of their language classes to complete the background knowledge questionnaires. The researcher was present for the duration of administration, to answer any questions (in Farsi or English). Overall, the questionnaire was found to be straightforward and easy to complete although its length was not always favorably viewed.

3.3.3.3 Data Processing

3.3.3.3.1 Spoken Data

A parallel forms reliability design required the administration of two forms of the speaking test to each participant resulting in 164 speaking tests (82 participant x 2 speaking tests). The focus of the study in comparing speaking performance across topics necessitates each topic (within each speaking test) to be rated separately. In Weir and Wu (2006), where a similar research design was adopted, the raters were asked to listen to the full speaking test and to give separate scores for each task. However, such an approach runs the risk of a halo effect, where raters are more likely assign similar scores for each topic when they are aware that they are marking the same person's performance. In order to address this potential problem, the speaking test for each participant was edited, divided into its constituent topics, anonymised and identified by different codes.

The editing process was carried out with Adobe Audition 1.5 (Adobe Systems, 2003-2004), a software package designed for editing audio files. The following steps were taken in the process:

- Playback of each speaking test in Adobe Audition
- Removing participant reference code from the beginning of each interview
- Marking the beginning and end of each task item, cutting the speech wave and pasting it into a new file
- Coding each new file using a system that identified the task, speaking form and participant reference codes. The 10 digit coding system was designed to make it difficult for raters to identify the same person's performance
- Once all files were prepared, the speaking files were divided into batches for marking (see section 3.4.3.3 for the rating design of the study).

3.3.3.3.2 Background Knowledge Questionnaire Responses

Participant responses to background knowledge questionnaires were entered into Excel and prepared for analysis.

3.3.3.3.3 C-test Responses

When piloting the C-test, the issue of C-test scoring in terms of semantic, syntactic and orthographically acceptable variants was examined. A decision was made to score the C-tests using two methods. First, using a dichotomous scoring method where fully correct answers were counted as correct while partially correct responses (either semantically or syntactically) were counted as incorrect. Minor spelling errors were counted as correct. Second, a partial-credit scoring method was adopted where partially acceptable semantic and syntactic variants were given a score of 1, fully correct responses were given a score of 2 and an incorrect response was given a score of 0. A partial-credit scoring system was

found to pertain predominantly to inflections such as third person 's', plural 's' or tense markers, but also to cases in which a choice of noun or adjective would fit in with the meaning of the context but would not be 100% grammatically or semantically accurate.

The argument for opting for a partial-credit scoring system is also closely linked to the measurement model which was planned to be used with the scores i.e. Rasch measurement. The unidimensionality assumption of the Rasch model holds that performance on test items should reflect the underlying abilities of the test takers on a single dimension. In scoring the C-tests, it was felt that partial knowledge of an item might indicate higher ability of a person on a language proficiency continuum than someone who had not attempted the item or answered it incorrectly. A drawback of such an approach is claimed to be a drop in the reliability of scoring and its objectivity. However, such objectivity in scoring C-tests may not always be desired as it might conceal answers that 'are possible but deviant' (Lee-Ellis, 2009:254). In order to address these concerns, it was decided to use both scoring systems and to examine which system exhibits better fit to the Rasch model (the results of both analyses are discussed in full detail in section 4.1).

3.4 Data Collection (Phase II)

3.4.1 Rater Selection Criteria

Given the use of IELTS Speaking Tests as the main instrument of data collection in this study, it would have been preferable to seek the assistance of trained and experienced IELTS examiners, as they would be familiar with the format of the test, the ratings scales and the rating process which would increase rating reliability and consistency. This option

was not available²². However, the literature review chapter demonstrated that rater training has very little effect on relative severity and leniency (which is the major concern of the study) and that even without training individuals can exhibit high consistency levels within themselves (see section 2.2). Moreover, given the adoption of MFRM for analysis, relative rater severity and leniency could be parameterised, thus obviating the need for having IELTS examiners. It was also believed that with some basic training, individuals with extensive teaching experience and from educational backgrounds in language teaching/testing-related fields would be able to complete the ratings to a sufficiently satisfactory level. To this end, an e-mail was circulated to current/former colleagues at Oxford University's Department of Education in which interested individuals were invited for collaboration. The purpose of the study, a description of what the rating process would entail and the preliminary rating schedule were included in the e-mail (see Appendix M).

Responses were screened and rater selection took place on the basis of the following criteria:

- English language teaching and/or examining experience
- Familiarity with language proficiency interviews e.g. IELTS, Cambridge Main Suite exams, TOEFL)
- Graduate degree in TESOL, applied linguistics, language assessment or a similar field
- Availability in the given time period

On the basis of these criteria, four raters were invited for participation in the study. Table 3-3 provides rater profiles in terms of gender, educational background and language teaching/testing experience.

²² IELTS examiners can collaborate in projects which have gained approval/permission from the IELTS partners.

Table 3-3 Rater Profiles

RATER	1	2	3	4
Gender	Female	Female	Male	Female
Education	Msc. in Applied Linguistics and Second Language Acquisition CELTA/DELTA holder Reading for a PhD in Education with an applied linguistics focus	Msc. in Applied Linguistics and Second Language Acquisition	Msc. in Applied Linguistics and Second Language Acquisition PhD holder in corpus linguistics	Msc. in Applied Linguistics and Second Language Acquisition Enrolled in a DPhil in Education with an applied linguistics/assessment focus
TESOL/ Testing experience	6+ years of TESOL experience (exam preparations, business English, young learners, IELTS preparations)	10+ years of TESOL experience Exam preparations for Cambridge Main Suite and IELTS Familiarity with rating criteria of language proficiency interviews	Extensive TESOL experience, CELTA/DELTA instructor Researcher and consultant in language assessment projects	Extensive TESOL experience, Familiarity with IELTS (focus of Masters dissertation)

3.4.2 Instruments

3.4.2.1 Rating Scales

The public version of the IELTS Speaking Band Descriptors (Appendix B) was used for scoring spoken performances. This analytic rating scale has been extensively used in high-stakes assessment contexts for over 10 years and is grounded in empirical research and performance sampling²³. It also corresponds directly to the IELTS Speaking Test tasks used in the study. Luoma (2004) invites researchers to tailor available rating scales for their own needs. In line with this recommendation, the IELTS rating scale was slightly

²³ Performance sampling allows experts to work with spoken samples that have been rated by a number of different raters. Levels of performance are subsequently negotiated on the basis of observed features and comparisons with the scale descriptors.

modified in order to reflect the design of the study (i.e. rating individual tasks as opposed to rating the whole speaking test).

One of the weaknesses of the IELTS rating scale identified in the literature (see section 2.1.5) and commented on by IELTS examiners (as documented in Brown and Taylor, 2006) was the lack of a content-oriented or a topic development criterion. From a theoretical perspective, Sato (2012) questions the reason why such (non-linguistic and content-oriented) criteria have not been defined as part of the construct definition of some general oral proficiency tests (such as the IELTS speaking test), arguing that ‘narrowly restricting our focus to linguistic features may lead to erroneous inferences about L2 learners’ ability to communicate effectively’ (237). In support of this argument, Sato (2012:223) provided empirical research which showed that the content of performance of Japanese students’ monologues on different topics, rated on a *content elaboration/development* scale, amongst other criteria, made a ‘substantive contribution’ to raters’ ‘intuitive judgments of oral proficiency’. The findings led the author to conclude that ‘the quality of the ideas that test-takers attempt to convey should be treated as a criterion in oral assessments’ (Sato, 2012:237).

In light of the above literature and given the focus of the study on the effects of topic on performance, a decision was made to include a content-oriented criterion, hypothesised to be a more direct reflection of the topic-related content of performances vs. more linguistic-related aspects of performance (in line with Sato, 2012). Having surveyed a number of holistic and analytic rating scales used e.g. from the Common European Framework of Reference (CEFR), Test of English as a Foreign Language (TOEFL), American Council on Teaching of Foreign Languages (ACTFL), the scale that best suited

the purpose of the current study was the *topic development* component of the Internet-Based TOEFL (TOEFL iBT) for independent tasks (ETS, 2004). This scale was used for in the study and can be accessed in Appendix N.

3.4.2.2 Observation Checklist

An observation checklist, based on the work of O’Sullivan, Weir and Saville (2002), was used in this study. This instrument was originally designed as an efficient and practical method for evaluating speaking test tasks in terms of the language functions that they elicit. It is also used for ‘real-time’ comparisons of the language functions elicited in different task types and/or parallel tasks:

[The checklist] enables both *a priori* and *a posteriori* analysis of speaking task output...[it allows] language samples elicited by the task to be scanned for... functions in real time, without resorting to the laborious and somewhat limited analysis of transcripts.

O’Sullivan, Weir and Saville (2002:33)

In the context of this study, such an instrument allows for the comparison of the different topics (within each task type), in terms of their relative capacity to elicit a range of functions. The original observation checklist by O’Sullivan, Weir and Saville (2002) was modified for use in this study²⁴ and includes the following functions: providing personal and non-personal information, expressing and justifying opinions, explaining, suggesting, expressing a preference, comparing, contrasting, paraphrasing, narrating, describing, elaborating, summarizing, paraphrasing, speculating, staging, analyzing, negotiation of meaning, conversation repair, self-repair, agreeing, disagreeing, commenting on task or topic difficulty and asking for clarification. The checklist was accompanied with an

²⁴ The original checklist was designed for speaking tests with a paired format. These tests are designed to elicit a broader range of interaction functions which might not occur in a language proficiency interview format, such as the one used in this study. In line with Brooks (2002; 2003), the original checklist was modified to incorporate functions that frequently occur in tests such as IELTS.

additional information sheet (based on O’Sullivan, Weir and Saville, 2002), giving further clarifications on some of these functions (see Appendix O).

3.4.2.3 Instructions/Score Sheets

An Excel-based score sheet (see Appendix P for a sample score sheet) was designed for raters, providing the format necessary for inserting ratings for the speaking tasks, recording instances of observing particular types of functions and any additional comments for each task code. Instructions for filling in the score sheet were separately provided (Appendix Q).

3.4.3 Rating Procedures

3.4.3.1 Allocating Speaking Tasks to Raters

A fully-crossed rating design where all raters rate all performances was not feasible in the current study due to the large number of speaking tasks. Instead, a ‘common batch’ model was used following Weir and Wu (2006:175):

The common batch ‘[can] offer a practical solution to one of the requirements of MFRM, namely that there should be a degree of overlap for each of the facets. In this case it meant that candidates had to be connected through an overlap in the tasks taken and in the raters marking the tapes.

The 809 speaking tasks in the study were divided into several²⁵ batches: one common batch – rated by all raters – and four additional batches, rated by individual raters. Each batch contained a random allocation of performances from all tasks (topics). Similar to Weir and Wu (2006), participants in the study are connected through an overlap in tasks (two common tasks of family and leisure time), while raters are connected through an

²⁵ The number of batches in this study was determined by the funding available for rater recruitment and compensation.

overlap in the tasks they rated (batch 1: the common batch). The rating design of the study is illustrated in Table 3-4.

Table 3-4 The ‘Common Batch’ Rating Design of the Study

RATER	COMMON BATCH (1)	BATCH (2)	BATCH (3)	BATCH (4)	BATCH (5)
Rater 1	X	X			
Rater 2	X		X		
Rater 3	X			X	
Rater 4	X				X

3.4.3.2 Rater Familiarisation Process

One-to-one meetings were held with each rater during which the raters were familiarised with the various instruments in the study (as discussed above). Basic training was provided in the form of an *IELTS Scores Explained* standards-setting DVD (UCLES, 2006) which they were asked to watch at their convenience. This DVD explains the rating of the IELTS Speaking Module and presents benchmark performances at each band level after which the scores for those performances are displayed. The raters were asked to first familiarise themselves with the rating scale, to watch the DVD and to assign scores to each performance and to subsequently check their scores against those awarded on the DVD and to gauge their marking accordingly.

Raters were also asked to familiarise themselves with the observation checklist and were presented with a copy of the O’Sullivan et al. (2002) article in order to further acquaint themselves with the instrument and its purpose.

3.4.3.3 Further Clarifications and Batch Assignment

A second meeting was held with each rater during which their rating experience with the performances on the *IELTS Scores Explained* DVD (UCLES, 2006) was discussed and clarifications were provided if requested. At the end of the second meeting, each rater received a folder that contained the following:

- A CD or USB stick containing the speaking tasks allocated to each rater (i.e. the common batch and an additional batch)
- An ‘instruction to raters’ sheet, which detailed the sequence of steps to be taken in the rating procedure and guidelines for dealing with corrupt or noisy files, etc.
- A hard copy of all instruments (an electronic copy was also e-mailed to the raters)

3.4.3.4 Post-rating Semi-structured Interviews

Once ratings were completed, a short interview was conducted with the raters in order to gain an insight into their rating experience. They were asked to comment on the rating process, the rating criteria, and any aspect of the rating process that they found to be problematic or challenging. They were also asked to comment on their observations regarding (any) topic/topic unfamiliarity effects on their rating decisions. It was hoped that these qualitative perspectives could complement the results of the quantitative analyses.

3.4.4 Data Processing

3.4.4.1 Speaking Scores Data Set

Once data were received from raters, the score sheets (speaking scores and functions) were screened for any potential problems (e.g. missing files, incorrect data entry, etc.). A data set was then built in which the ratings of all tasks for the same participant were

identified and brought together in the following structure: [Person Facet ID, Rater Code, Topic Code, Rating Criteria]

The data set was constructed in SPSS 19.0 (IBM, 2010) and subsequently transferred into fixed text formats for use with FACETS (Linacre, 2011b) and RUMM 2030 (Andrich et al., 2010).

3.4.4.2 Interview Transcription

The interviews with raters were transcribed and prepared for subsequent qualitative analysis.

3.5 Quantitative Data Analysis

3.5.1 The Rasch Models

The Rasch family of models was used as the main method of analysis for a) examining the quality and functioning of the different instruments used in the study and b) the construction of linear interval-level measures of examinee spoken proficiency, background knowledge and general language proficiency.

3.5.1.1 Background Knowledge Questionnaires

Participant responses to the 5-point Likert scale questionnaire items were analysed with the partial-credit Rasch model and using RUMM 2030 (Andrich et al., 2010). The fit of the questions to the Rasch model were investigated. Linear measures of participants' background knowledge of topics were constructed (see Chapter 4). These measures were subsequently used for grouping purposes in a 5-Facet MFRM. The linear measures were also used as a predictor variable in a multiple-regression analysis of spoken score data.

3.5.1.2 C-tests

The C-test scores from the two dichotomous and partial-credit marking systems were analysed with the dichotomous and polytomous Rasch models respectively. The results of the dichotomous marking system suggested a better fit to the Rasch model which is why these scores were retained for further analysis. The blanks for each text in the C-test were summed and each text was analysed as a ‘super-item’ as treating the blanks in each text as separate items is a violation of the assumption of local independence (see Chapter 4 for a full description of all analyses). The resultant linear measures were used for grouping purposes in subsequent MFRM bias analyses. The measures were also directly used as a predictor variable in a multiple regression analysis of spoken score data.

3.5.1.3 Speaking Score Data

The speaking performance scores from each task topic were analysed using Multi-Faceted Rasch Measurement (Linacre, 1989). First, a four-facet MFRM was carried out with examinees, raters, topics and criteria as the facets. Second, a five-facet MFRM was carried out where background knowledge of topics was conceptualised as an additional facet. Third, a series of bias analyses were carried out in MFRM. These analyses were subsequently drawn upon to answer the majority of the study’s research questions.

3.5.2 Multiple Regression Analysis

The linear background knowledge and general language proficiency measures, in addition to task type and gender were used as predictor variables for topic-based spoken performance ability measures of participants (using a 3-Facet Rasch model) in a multiple regression analysis using SPSS 19.0 (IBM, 2010).

3.5.3 Frequency Statistics

The results of the functions observation checklist were analysed using descriptive frequency statistics in SPSS 19.0 (IBM, 2010).

3.6 Qualitative Data Analysis

The table of unexpected responses from the 4-Facet MFRM analysis and the results of the background knowledge measures were used as a selection criterion for speaking performance scores where lack of background knowledge was hypothesised to exert maximum influence. The selected performances were subsequently transcribed and the content of the performances were analysed qualitatively. These transcripts were repeatedly scanned for the purpose of code generation (in line with Ellis and Barkhuizen, 2005). Next, recurrent patterns and themes were identified and subsequently used in providing a different perspective on the complex relations between topics and test takers' interaction with topics.

A similar qualitative process was applied to the rater interview transcripts.

3.7 Summary

This chapter has summarised the methods of data collection and analyses used in the study. Schematic representations of the different stages of data collection and analyses as well as the various instruments used at each stage are provided in Figure 3-4 and Figure 3-5 respectively.

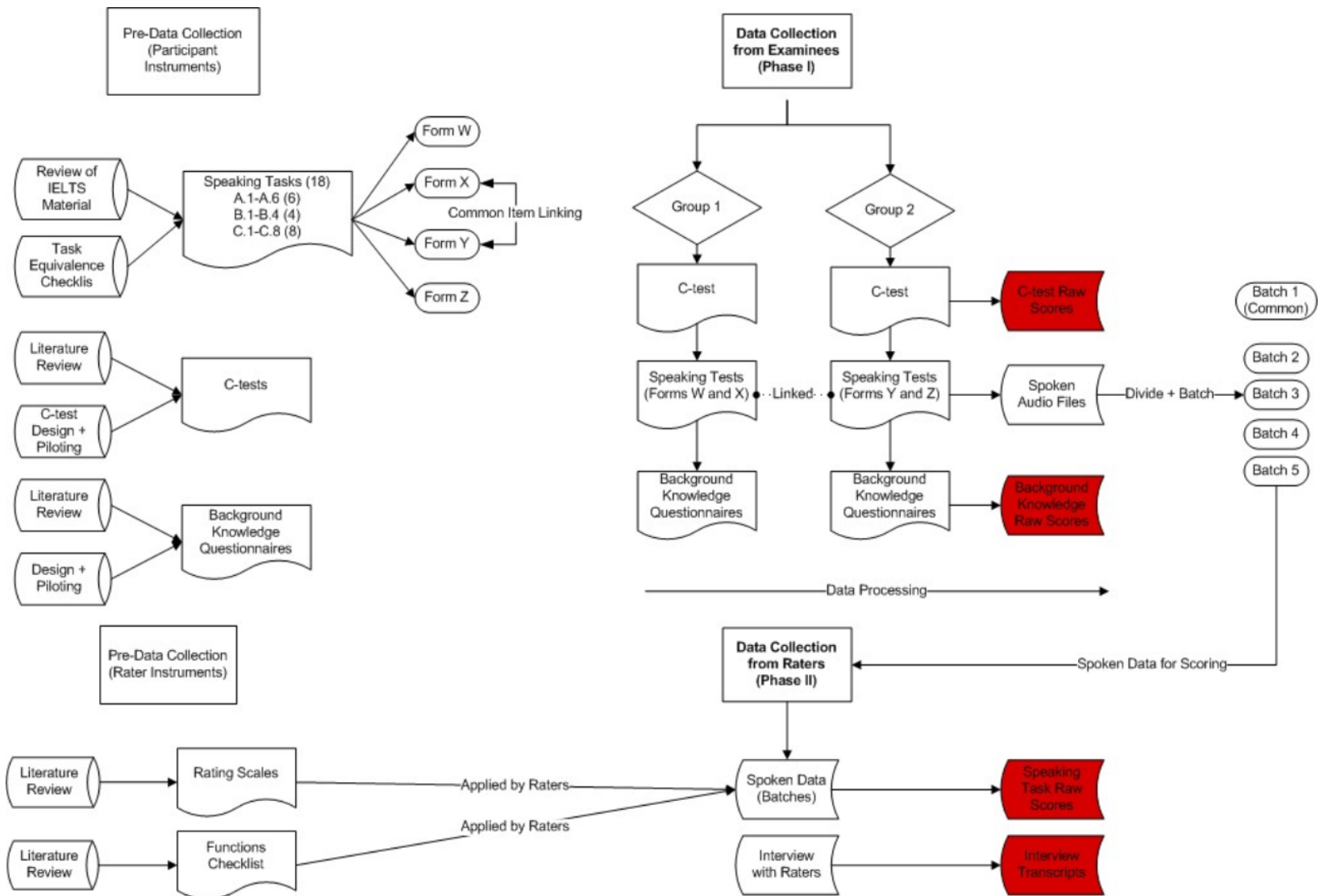


Figure 3-4 Data Collection Process

Data Analyses Procedures

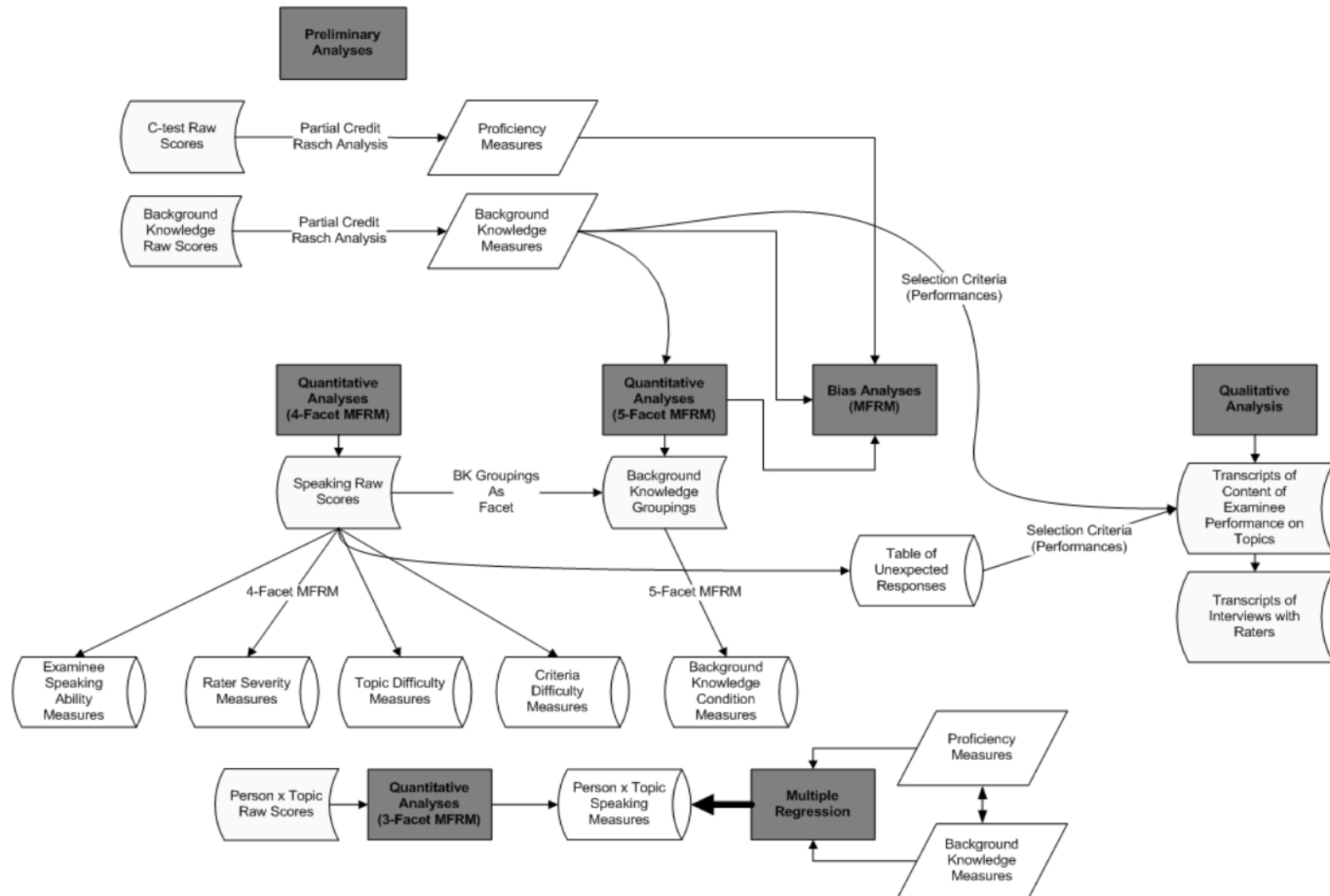


Figure 3-5 Data Analysis Procedures

Chapter 4 Preliminary Analyses

This chapter reports the results of C-test pilot study as well as the preliminary analyses of participant responses to the C-tests and Background Knowledge (BK) questionnaires. These results are subsequently used for the study's main analyses of speaking score data (Chapter 5). Moreover, in examining the psychometric properties of these instruments (using Rasch analysis), the findings from this chapter provide an insight into the quality and functionality of the two instruments.

4.1 C-tests

In this section, the results of the C-test pilot study are first discussed before reporting the C-test findings from the main data collection phase. An overview of C-tests was provided in the previous chapter (see section 3.3.2.3) in which a case was made for using C-tests as a measure of general language proficiency.

4.1.1 C-test Pilot Study: Overview

Following an initial pilot study with native speakers of English, the study's C-tests (see Appendix I) were piloted with a convenience sample of 203 German speakers of English as a second language. The pilot study served two main purposes:

1. Informing the C-test scoring system to be adopted for the main study
2. Evaluating the functionality of the C-test as a measure of general language proficiency in distinguishing between persons from different proficiency levels

In addition, the administration of the C-tests in the pilot stage helped determine a practical time frame for completion of the tests while the scoring of the C-tests allowed for modifications to the answer key in specifying acceptable variants.

4.1.2 C-test Pilot Study: Analysis using the Rasch model

As discussed in the methodology chapter (see section 3.5.1.2), the C-tests were marked using two different scoring systems: dichotomous and partial credit. These scores were subsequently analysed using the dichotomous and polytomous Rasch models respectively²⁶. The latter model is used for items with more than two response categories (Tennant and Conaghan, 2007). The reason why the Rasch model is relevant to this analysis is succinctly captured in the following quote:

Rasch analysis allows for a unified approach to several measurement issues, all of which are required for the validity of the transformation to interval scaling: testing the internal construct validity of the scale for unidimensionality, required for a valid summed raw (ordinal) score; testing the invariance of items (that is, the ratio of difficulties between any pair of items remains constant across the ability levels of respondents), required for interval-level scaling; appropriate category ordering (whether or not the category ordering of polytomous items is working as expected); and differential item functioning (DIF; whether bias exists for an item among subgroups in the sample).

Tennant and Conaghan (2007:1359)

The scores from items (blanks) on a C-test are designed to be summed in order to provide a measure of general language proficiency. Rasch analysis can be used to ensure that these raw scores can be transformed into an interval-level scale and that the items reflect an underlying unidimensional trait of general language proficiency. Moreover, the analysis can provide ‘stable estimations of examinee ability and item difficulty’ (Lee-Ellis, 2009:

²⁶ The dichotomous and polytomous Rasch models refer to the mathematical derivation of the model: ‘When items in a scale have only two response options, the dichotomous model is chosen. When items have 3 or more options, then the Rasch model takes a slightly different form, either what is commonly known as the Andrich Rating Scale Model (Andrich, 1978) or the Masters Partial Credit Model (Masters, 1982). These derivations use the same Rasch model, but they differ slightly in their mathematics and are named after the people who developed them. The principal difference between the two is that the former expects the distance between thresholds to be equal across items (Tennant and Conaghan, 2007:1359).

254) while also allowing the means for evaluating the quality of the scale in the form of fit statistics (Tennant and Conaghan, 2007; Bond and Fox, 2007).

In this thesis, two different software packages are used for running Rasch analysis: RUMM 2030 (Andrich et al., 2010) for dichotomous and polytomous Rasch analyses and FACETS (Linacre, 2011b) for Multi-Faceted Rasch Measurement. It should be noted that the statistics reported in different software packages are not directly comparable (Tennant and Conaghan, 2007). In RUMM 2030 (Andrich et al., 2010), the difference between the observed values and those expected by the Rasch model are labeled chi-squared statistics. The output also reports a fit residual statistic which is based on the standardised residuals of all the responses of all the persons to an item (or criterion). It is by considering these different pieces of information, amongst others, that the extent of data fit to the model can be evaluated.

4.1.3 C-test Pilot Study: Item Analysis

C-test scores from the dichotomous and partial-credit scoring methods were analysed using RUMM 2030 (Andrich et al., 2010). In these analyses, each blank in the C-test texts was treated as an independent item. Across all three versions of the C-test, there were 279 items in total, linked through items on a common text (common item linking), with responses from 203 persons. The overall fit of the resultant scores (from the two scoring methods) to the Rasch model is examined by considering a number of different fit statistics: the overall item and person fit statistics, reliability indices and the percentage of misfitting persons and items (see Table 4-1 for the summary statistics).

In RUMM 2030, the overall fit of the data to the model can be evaluated by first considering the overall item fit statistic, which is a statistic that provides information on

the ‘fit of the data to the model from the perspective of the items’ (RUMM laboratory, 2009:5). Given that this statistic approximates a standard normal deviate, it should be interpreted under the hypothesis that if the data fit the model, then the deviations between responses and the model are attributable only to random errors in which case the mean of the items (M) would be close to 0 and the standard deviation (SD) close to 1. A similar interpretation is applicable to the overall person fit residual statistics with expected mean and standard deviation values of 0 and 1 respectively. Table 4-1 shows that the residual mean and standard deviation values for items are relatively close to their expected values across the two scoring methods with the partial-credit scoring method exhibiting a mean value (M=-0.09) slightly closer to the expected value of 0.00 compared to the dichotomous model (M=-0.014). The overall person fit residual statistics are similar across the two scoring methods, both exhibiting deviations from their expected values.

Table 4-1 Summary Statistics for the Analyses Based on the Two Scoring Methods

SUMMARY STATISTICS	SCORING METHOD	
	Dichotomous	Partial-Credit
Fit Residual (Items)	Mean=-0.14 SD=0.96	Mean=-0.09 SD=0.89
Fit Residual (Persons)	Mean=-0.19 SD=0.72	Mean=-0.17 SD=0.75
Total Item-Trait Interaction	$\chi^2=1208.6$ $df= 1040$ $p=0.00$	$\chi^2=1244.7$ $df= 1040$ $p=0.00$
Person Separation Index	0.92	0.92
% of Misfitting Items (fit values > 2.5)	5.4	5.7
% of Misfitting Persons (fit values > 2.5)	0.3	1.0

N=203, SD= Standard Deviation, χ^2 = Chi-Squared, df = degrees of freedom, p =probability

Fit of the data to the Rasch model can also be examined by considering another index: the total item trait interaction statistic which is the sum of individual item chi-square statistics. A large chi-square value with a significant probability value indicates misfit: 'a significant chi-square indicates that the hierarchical ordering of the items varies across the trait, compromising the required property of invariance' (Tennant and Conaghan, 2007: 1360). The results in Table 4-1 show significant misfit of the C-test score data to the Rasch model on the basis of the chi square item-trait interaction and highly significant probability values for both scoring methods. This finding is not surprising: the format of the C-test is in itself a violation of the assumption of local independence in the Rasch model, as the response to one item can influence the response to subsequent item(s). In dealing with this problem and in removing response dependence, a common practice is to sum the dependent responses and to treat each text in the C-test as a super item (Lee-Ellis, 2009). This is the focus of section 4.1.4. For now, a decision still needs to be made on the data set to be used (i.e. from the dichotomous or partial credit scoring methods).

A consideration of the remaining indices in Table 4-1 shows close similarities between the overall results of the two scoring methods. For example, the Person Separation Index (PSI), which in RUMM 2030 is an indication of whether the persons are spread out across the underlying continuum, is the same across both data sets. Generally speaking, a high value of close to 1 is preferred for this index. The high value of 0.92 shows that regardless of the scoring method adopted, the C-test can reliably distinguish between persons from different language ability levels.

The final values to consider in the table are the percentage of misfitting items and persons. In RUMM 2030, fit residual values that fall within the range of -2.5 to +2.5 are considered

as fitting the model. Table 4-1 shows that despite small percentages of misfit, the partial-credit scoring method has resulted in a larger percentage of misfitting persons and items. One possible explanation for the higher number of misfitting persons and items for the partial-credit scoring method is that what has been scored as showing partial knowledge may in fact reflect random guessing and thus yielding some unreliable results.

The summary statistics for both scoring methods display very similar results and do not show any clear advantages for either approach. We therefore turn our attention to a consideration of thresholds and category functioning of the polytomous items in evaluating the partial-credit scoring method. For polytomous items ‘whether the responses to the items are consistent with the metric estimate of the underlying construct is indicated by an ordered set of response thresholds for each of the items’ (Tennant and Conaghan, 2007: 1360).

The threshold map and category probability curves for the 279 items from the polytomous Partial-Credit Rasch model (Masters, 1982) were carefully studied²⁷. Results showed that the partial-credit scoring system resulted in items which displayed reversed thresholds for *all* except 28 items. Likewise, the categories were not functioning as intended. For illustration purposes, three representative examples of such items are selected and their threshold and category probability curves are reproduced in Figure 4-1.

The threshold probability curves (TPCs) all display disordered thresholds, as threshold 1 (in blue), which is designed to divide the easier response categories of 0 and 1, is to the right of threshold 2, which is designed to divide the more difficult response categories of 1 and 2, whereas the opposite should be the case. The categories are therefore not

²⁷ There were too many figures to be reproduced.

functioning in accordance to the implicit idea that ‘the thresholds between higher level categories are more difficult than thresholds between lower level categories’ (Van Wyke and Andrich, 2006:21). The category probability curves (CPCs) for the selected items also tell a similar story. As it can be seen, there is no region of the ability continuum in which category score of 1 is most likely to be achieved. As the probability of getting a score of zero decreases, it is the likelihood of achieving a score of 2 that increases, suggesting that the categories are not working as intended. Furthermore, in the region of person locations where one would expect a category 1, persons are more likely to receive either a 0 or 2 that is an incorrect or fully correct answer; in other words, a dichotomous response.

Evidence from the disordered thresholds and lack of category functioning of an overwhelming majority of the partially-scored items, provides a strong case for opting for the dichotomous scoring method over the partial-credit method. Moreover, even when remedial action was taken by collapsing categories, there were no substantial improvements in the fit of the data to the model. The remedial action involved the following steps:

- Dysfunctional categories were collapsed by rescoring the problematic items
- The 28 items that were working properly in terms of threshold ordering and category functioning were retained
- The analysis was re-run with the rescored items and the original 28 items
- Results showed very little improvement to the fit of the data to the model with a slight 0.2 increase in the reliability index. For this reason, a decision was made to opt for the more simple dichotomous scoring method, which also yielded a lower percentage of misfitting persons and items.

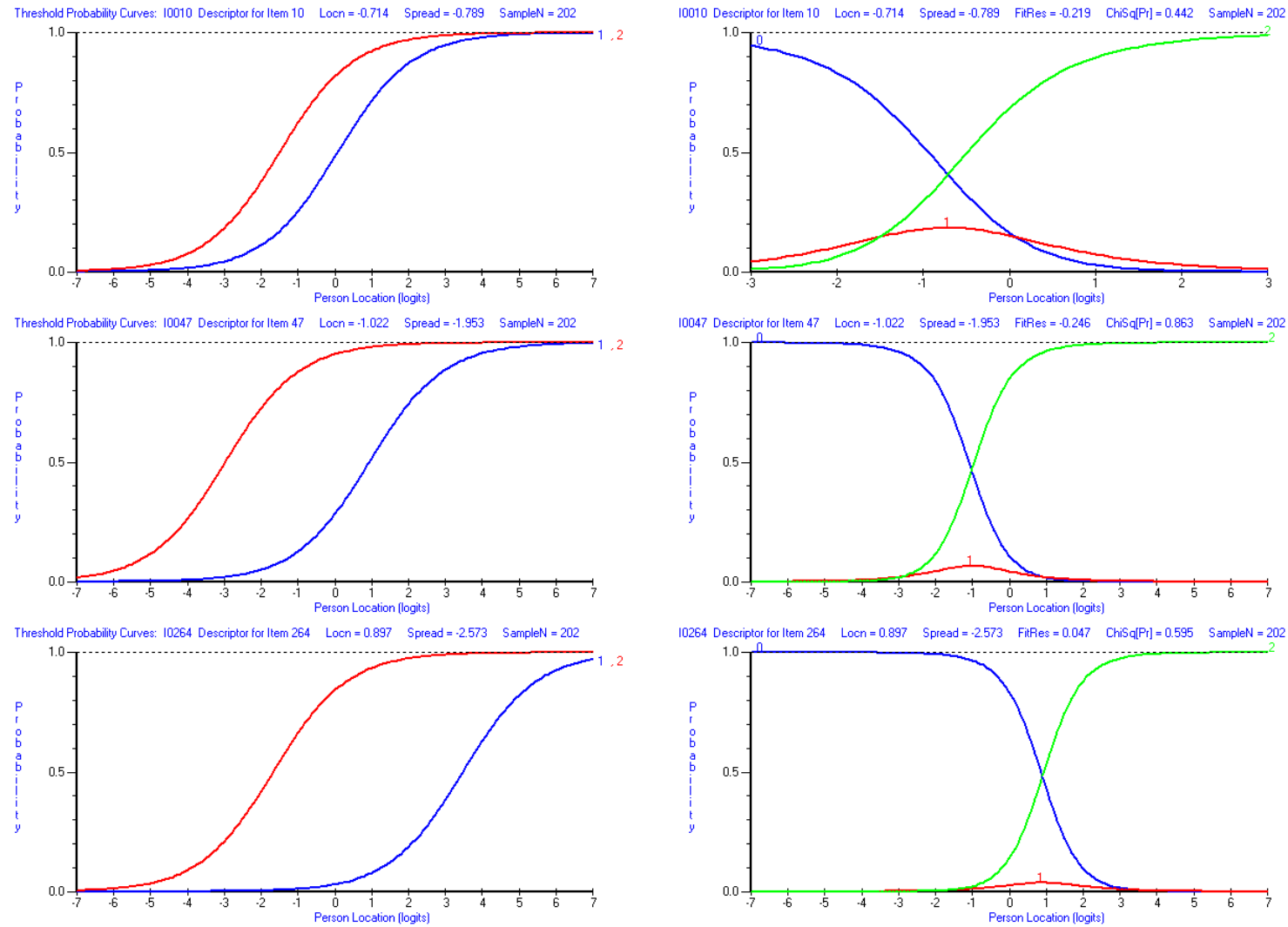


Figure 4-1 Threshold Probability Curves (right) and Category Probability Curves (left) for Items 10, 47 and 264

4.1.4 C-test Pilot Study: Super-item Analysis

4.1.4.1 Violation of the Assumption of Response Independence in C-tests

The main assumption of the Rasch model is that of *independence of responses* or *local independence* which is the notion that any variation in responses of persons to an item should only reflect the variable (or the ability) under study and that ‘for the same [ability] value, there is no further relationship among responses’ (Marais and Andrich, 2008:201). Marais and Andrich (2008) make an explicit distinction between the two ways in which this assumption can be violated. On the one hand, variables (or parameters) other than the ability under investigation may be influencing responses, in which case the assumption of ‘*unidimensionality*’ has been violated. In other words there is ‘*trait dependence*’ (Marais and Andrich, 2008:281). The other type of violation termed as ‘*response dependence*’ (*Ibid.*) refers to the dependence of the response of one item to the previous item(s). The relevance of the latter violation in respect to C-tests is very clear; a correct answer in one blank can provide more contextual clues for filling subsequent blanks. Andrich (1985:256) brings forward both practical and theoretical reasons for treating the items that belong together as a ‘*subtest*’ (also referred to as super-item in the context of C-tests); where each subtest represents a unit ‘and can be characterized by two parameters...and the parameterization is clearly more parsimonious than if each item was parameterized separately. This parsimony should make the parameter estimates of a subtest more stable than the parameter estimates of each item’.

In line with Andrich (1985) and Marais and Andrich (2008), a *subtest* or henceforth super-item analysis was carried out with the C-test data, where dichotomously scored items (blanks) that belonged to the same short texts were combined to build higher-order polytomous items and then reanalysed. In other words, the 275 items, which were treated independently in the previous section, were *bundled* (Rosenbaum, 1988) into the 10 texts that they originally belonged to. The presence of response dependence amongst items manifests itself in a drop in the Person-Separation Index when items are summed, compared to when they are analysed independently (Marais and Andrich, 2011).

4.1.4.2 Results of C-test Super-Item Analysis

Responses to the dichotomously scored C-tests were summed and subsequently fitted to the Rasch model using RUMM 2030 (Andrich et al., 2010). The partial-credit parameterization of the Rasch model for polytomous data (Masters, 1982) was selected, as this parameterization makes no assumptions regarding the equivalence of distances between thresholds in different categories, as opposed to the rating scale parameterization (Andrich, 1978), in which the distances between thresholds are held to be equal. An examination of the threshold map (Figure 4-2) suggests that the distances between thresholds are not equal which justifies the selection of the partial-credit parameterization. The map shows no reversed thresholds.

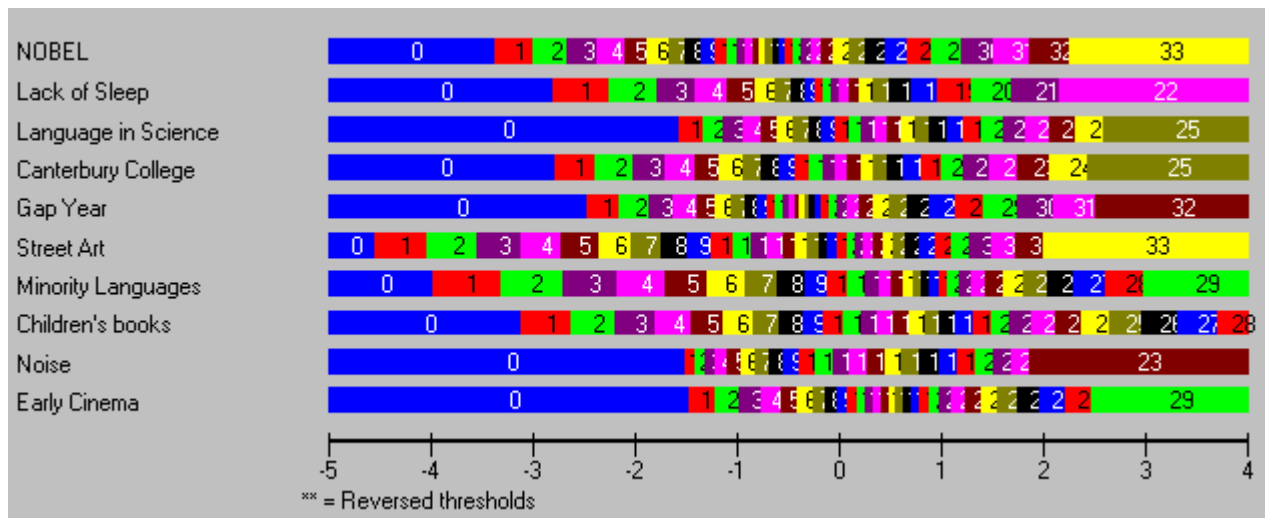


Figure 4-2 Threshold Map for the Texts (Super-items) in the C-tests

Similar to the previous section, the fit of dichotomously scored, super-item C-test data to the Rasch model was evaluated by considering a range of fit indices. The results of the overall item fit residual statistic ($M=0.05$; $SD=0.98$) show values which are very close to the expected values of 0 and 1 respectively. The fit residual values for the persons ($M=-0.27$, $SD=0.94$) are also very close to their expected values of 0 and 1. The chi-square item-trait interaction value ($\chi^2 = 43.10$, $df=40$) and its associated non-significant probability value of $p=0.34 > 0.05$ lend further support to the fit of the data to the model. As expected, the PSI in the super-item analysis had dropped to 0.90 (compared to 0.92 in the independent item analysis) which substantiates the presence of response-dependence in the original data. The new PSI reliability value (0.90) is still very high and indicates that the C-test can reliably distinguish between persons of varying underlying language abilities.

The fit residual statistics and chi-square probabilities of individual items were also examined which, in addition to the visual inspection of the Item Characteristic Curves (ICCs), suggest that the data fit the Rasch model very well. The ICCs for all super-items

are divided into the five easiest and five most difficult texts and reproduced in Figure 4-3 and Figure 4-4. These figures illustrate that the observations (in dots) are very close to their expected theoretical curves. In terms of person fit statistics, of the 203 pilot participants, only eight persons (less than 3%) had fit residuals which did not fall within the acceptable range of -2.5 to +2.5.

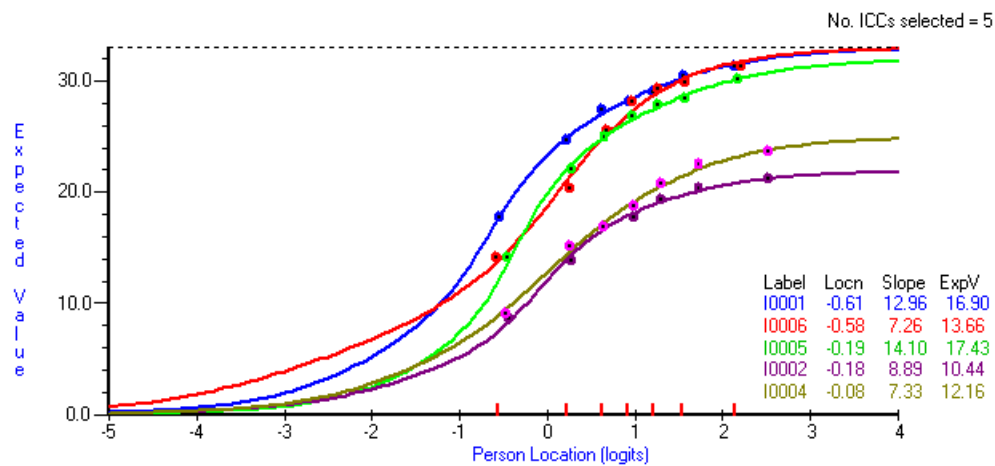


Figure 4-3 Item Characteristic Curves (ICCs) for the Five Easiest Texts (Super-items)

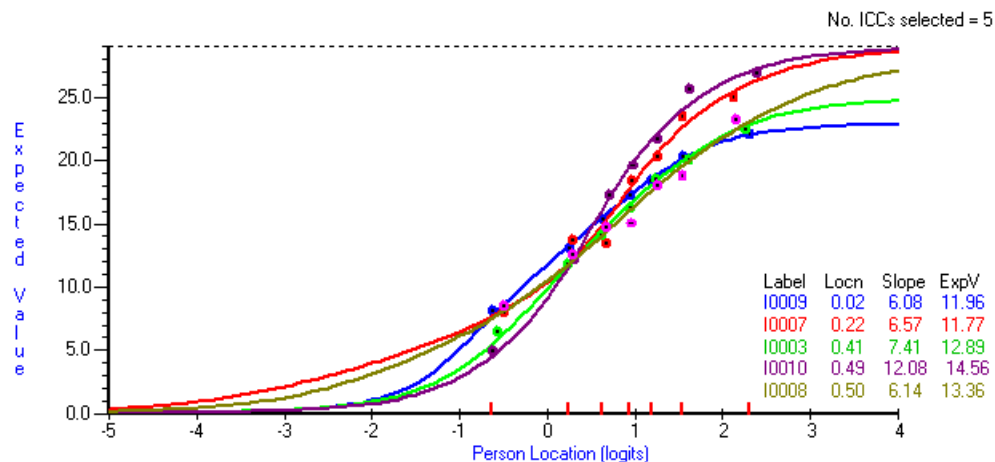


Figure 4-4 Item Characteristic Curves (ICCs) for the Five Most Difficult Texts (Super-items)

The final issue to be addressed is that of *targeting* of the test, which refers to the appropriateness of the measure(s) used for the population being assessed. Targeting can be evaluated by considering the mean location score obtained for persons against the item mean location of zero²⁸. If the value of the person locations is close to the item mean of zero, then the measure is neither too easy nor too difficult for the persons taking the test. Targeting can also be visually depicted in the person-item distribution map (see Figure 4-5).

The mean of the pilot participants (N=203) is $M=0.91$, $SD=0.65$, which is higher than the mean of the items. This relatively large mean value suggests that the test was easy for the pilot participants. This is not unexpected; the participants' self-reports of proficiency indicated that they were mostly upper-intermediate to advanced second language learners of English. Nevertheless, the results of these analyses and the excellent fit of the C-test data to the Rasch model substantiates the use of the C-tests as a reliable measure of general language proficiency which allows for the transformation of raw scores to estimated ability measures on an interval level logit scale. The C-tests were therefore used as a measure of general language proficiency in the main study.

²⁸ The mean of items in the Rasch model is arbitrarily centred at zero.

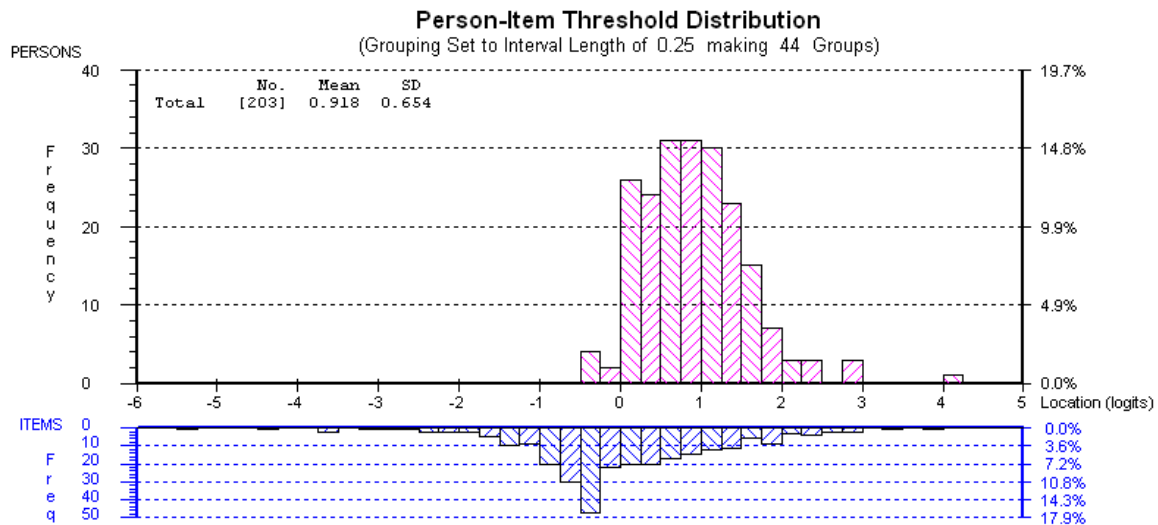


Figure 4-5 Person-Item distribution Map for the Super-item Analysis

4.1.5 C-test Results (Main Study)

The results of the pilot study supported the use of the designed C-tests as a measure of general language proficiency for the study. The C-tests were therefore administered to the 82 participants in the main data collection phase. Following the findings from the pilot study, responses were dichotomously scored. Each text in the C-test was subsequently treated as a super-item where the responses to individual items within each text were summed. The Partial Credit derivation of the Rasch model for polytomous items (Masters, 1982) was used to analyse the data using RUMM 2030 (Andrich et al., 2010).

The summary fit statistics from the main C-test analysis are presented in Table 4-2. The overall fit residual statistics for the items and the persons are generally acceptable and close to their expected values of 0 and 1. The total item trait interaction chi-squared value ($X^2=31.1$) and the corresponding non-significant probability value ($p=0.83>0.05$) suggest that the required property of invariance has been met. Moreover, the high person

separation index of 0.96 indicates that the C-test can reliably separate persons from different ability levels. Lastly, the table shows that there were no individual persons or super-items with fit values outside of the acceptable range of -2.5 to +2.5. Taken together, these results suggest an excellent fit of the participant data to the Rasch model.

Table 4-2 Summary Statistics for the Main C-test Data Analysis

SUMMARY STATISTICS	OVERALL FIT RESIDUALS (ITEMS)	OVERALL FIT RESIDUALS (PERSONS)	TOTAL ITEM TRAIT INTERACTION	PERSON SEPARATION INDEX	% OF MISFITTING ITEMS /PERSONS (FIT VALUES > 2.5)
	Mean=0.2 SD=0.86	Mean=-0.18 SD=0.83	$\chi^2=31.1$ $df= 40$ $p=0.83$	0.96	-

N=82, SD= Standard Deviation, χ^2 = Chi-Squared, df = degrees of freedom, p =probability

The super-item difficulty measures and fit statistics are produced in Table 4-3 where items have been arranged in order of increasing difficulty. Results show that the super items have a range of difficulty measures from -0.84 (for super-item/text no. 6) to 0.55 (super-item/text no. 10) covering a range of 1.39 logits. As illustrated, the fit residuals all fall within the acceptable range of -2.5 to +2.5. The chi-squared probability values are also all non-significant.

Table 4-3 Super-item Fit Statistics

SUPER-ITEM	LOCATION	S.E	FIT RESIDUAL	χ^2	DF	PROB
6	-0.84	0.05	0.76	3.237	4	0.52
1	-0.79	0.04	-0.647	2.419	4	0.66
5	-0.19	0.04	0.14	3.459	4	0.48
2	-0.16	0.07	0.047	3.425	4	0.49
4	0.01	0.08	-0.175	2.607	4	0.63
9	0.19	0.07	0.48	2.113	4	0.71
7	0.20	0.07	-0.124	3.482	4	0.48
8	0.49	0.08	1.472	1.803	4	0.77
3	0.53	0.05	0.566	7.504	4	0.11
10	0.55	0.07	-0.451	1.137	4	0.89

N (Items)=10, S.E= Standard Error, χ^2 = Chi-Squared, df = degrees of freedom, p =probability

The RUMM 2030 output for the person ability measures showed a range of person abilities from -2.18 logits to +3.80 spanning approximately 6 logits, indicating that the C-tests were successful in eliciting a range of language proficiency measures. The fit residual values of all persons fell within the acceptable range.

An examination of the Item Characteristic Curves, Category and Threshold Probability Curves did not reveal any unexpected patterns or disordered thresholds. All categories were also ordered, serving as further evidence for the quality of the C-tests used in this study as a measure of general language proficiency.

Finally a consideration of the targeting of the C-tests by examining the person-item (threshold) map (see Figure 4-6) indicates that the test items are well-targeted for the Iranian participants in the main study judging by the distribution of the item threshold difficulties relative to person ability measures. The mean of the persons ($M=0.25$, $SD=1.01$) is also very close to the mean of the items (0.00) showing that the test is neither too easy nor too difficult for the participants in this study.

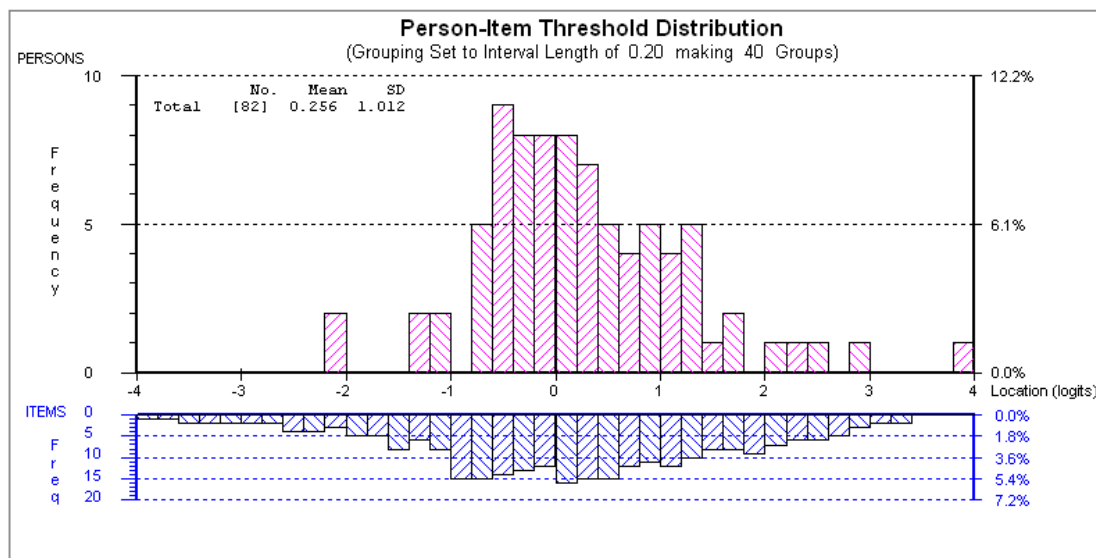


Figure 4-6 Person-Item Threshold Distribution Map (C-test Main Analysis)

In conclusion, the excellent fit of the C-test data to the Rasch model allows for the transformation of the participants' raw scores to estimated ability measures on an interval level logit scale. It is the general language proficiency measures from this analysis which are used in the main data analysis in Chapter 5. What is also relevant to this discussion is that when these general language proficiency measures were plotted against the participants' speaking ability measures from the MFRM analysis of speaking score data (to be discussed in full detail in Chapter 5), there was evidence of a strong positive correlation between these two variables ($r=0.91$, $p<0.001$), explaining 82% of the variance (see Figure 4-7). As explained in the methodology chapter (section 3.3.2.3), one of the reasons why C-tests were selected as a measure of general language proficiency in this study was on the basis of previous empirical evidence that suggested a strong correlation between such tests and tests of spoken language. The results from this study strengthen previous empirical findings and substantiate the use of the designed C-tests for the study's purposes.

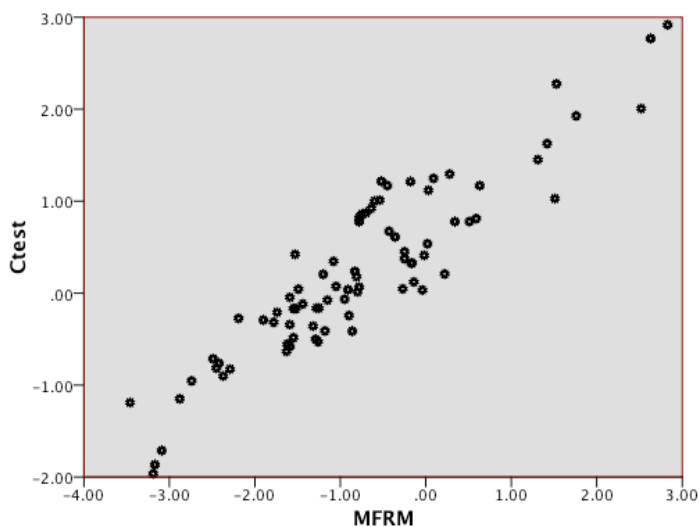


Figure 4-7 Scatterplot: Speaking Ability Measures vs. C-test Proficiency Measures

4.2 Background Knowledge Questionnaires

4.2.1 An Overview of the Questions, Scoring and Initial Observations

The background knowledge questionnaire (discussed in detail in 3.3.2.2) included eight questions on a five-point Likert scale that were repeated for every topic that the test takers had taken. These eight questions, which are hypothesised to elicit a measure of background knowledge of topics, are repeated below:

1. This topic is familiar to me.
2. The questions about this topic were easy to respond to.
3. I know *A LOT* about this topic, i.e., I have *MORE THAN ENOUGH IDEAS* to talk about this topic.
4. It was easy for me to produce enough ideas for this topic from memory.
5. If I were to talk about this topic in Farsi, I would have *MORE IDEAS* to talk about.
6. I had appropriate words to express my ideas about this topic easily.
7. I thought this was an interesting topic.
8. I performed very well on this task.

Scoring the questionnaires followed the conventional assignment of integer numbers in increasing order to the five options from Strongly Disagree (0) to Strongly Agree (4). All questions were positively worded, so the scores did not have to be reversed for any questions.

The scoring procedure provided some interesting insights into the way the respondents interacted with the questions. On the one hand, most participants displayed a particular pattern in answering the questions, and that pattern was more or less repeated for the topics that they found familiar or perhaps not particularly problematic. However, there was a break in the pattern of responses for those topics that the respondent had found unfamiliar or problematic, with the responses shifting from the *agree* end of the agree-disagree continuum to the *disagree* end. It is this striking shift in the pattern of responses

that can direct focus to fragments of performance where lack of background knowledge is mostly likely to manifest itself.

Question 5 elicited distinct patterns of responses. Most respondents would opt for the *agree* or *strongly agree* option regardless of the topic. However, the opposite was observed for a few participants. When crosschecked against their self-reports of proficiency, the latter had all reported themselves as high proficiency. It therefore appeared that the respondents' proficiency level was interacting with the way they answered the questions. This observation later informed part of the quantitative analysis of the questionnaire data.

Another interesting observation pertained to Question 7. It was hypothesised that having a lot of ideas about a topic would have a positive correlation with also finding the topic interesting. However, the initial scoring provided counter examples where respondents who rated a topic as unfamiliar reported having an interest in the topic and vice versa. It appeared that in the majority of cases, there was a positive relationship between the two. However, the pattern was not observed across all respondents.

Lastly, Question 8 required respondents to assess their own performance on a given topic. While some respondents provided a self-rating, there were a high proportion of respondents who opted for the *undecided* option.

It should be reiterated at this point that the data from the questionnaires were gathered for the purpose of quantitative analysis and for providing a measure of respondents' background knowledge of topics. Nevertheless, the above observations, which are qualitative in nature, can provide an insight into understanding the results of the quantitative analyses and even inform subsequent analyses.

4.2.2 Measurement and Likert Scales

Likert scales are prevalent in research in the social sciences. While they are regarded as a soft form of data collection and are highly subjective in nature, their resultant scores are summed and analysed without taking this subjectivity into account (Bond and Fox, 2007). The summing of scores from a questionnaire with Likert scales implies ‘that the additive structure of the data has been demonstrated’ and that the data are presumed to be interval (Bond and Fox, 2007:102). Consider the following two questions which are relevant to this study:

- a. I think that the choice of topic is not important for my final score.
- b. I think that the choice of topic is not important for my final score, if my English is good enough.

While both questions are eliciting the respondents’ attitudes towards the influence of topic on their scores, one can see that endorsing *disagree* for statement (b) might require much more of the underlying attitude, than endorsing *disagree* for statement (a).

Now consider examples of test-related anxiety statements:

- a. I am afraid of making grammatical mistakes in my speaking test.
- b. I am so afraid of making grammatical mistakes in my speaking test that I prefer not to take the test.

Once again, it is easy to see how the two stems may differ in eliciting varying degrees of attitude. However, in most traditional analyses of questionnaires, responses to the same option on the Likert scale for both stems are given the same raw score and are treated as equivalent, which according to Bond and Fox (2007:101) is ‘both counterintuitive and mathematically inappropriate’. The authors then suggest adopting the Rasch model to the

analysis of questionnaire data: ‘the Rasch model allows the item difficulty of each stem or question to be based on the way in which an appropriate group of subjects actually responded to that stem in practice’ (2007:103).

For the reasons outlined above, the questionnaire data were analysed using the Rasch model and the results are discussed in the next section.

4.2.3 Background Knowledge Questionnaire Results

The response data from the set of eight questions (polytomous items with five categories each) were fitted to the Rasch model using RUMM 2030 (Andrich et al., 2010), in order to assess the fit of the data to the model and to review the psychometric properties of the scale.

Each data line consisted of a participant’s responses to the eight questions on each topic. A total of ten sets of responses were therefore available for each participant, corresponding to the topics they had taken. Each person estimate therefore reflects a person’s measure of background knowledge on a specific topic. Topics were specified as a person factor, allowing for the background knowledge mean to be calculated for individual topics.

The partial-credit parameterisation of the model (Masters, 1982) was opted for, on the basis of evidence from the threshold map (Figure 4-8), which demonstrates that the distances between the response categories for the different question items are not equal. The figure also exhibits reversal of thresholds for questions 1 and 5 (marked with **), which will be addressed separately in the next sections.

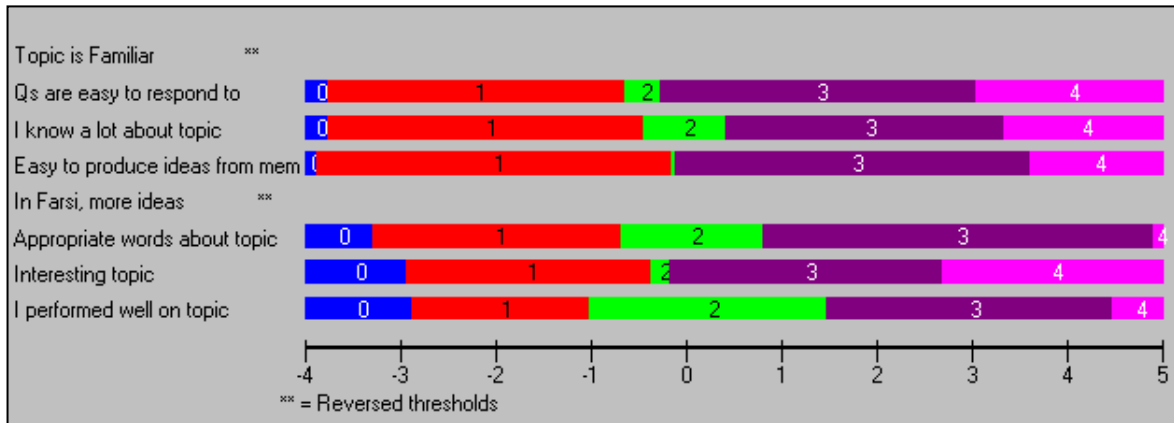


Figure 4-8 Threshold Map for the 8-item Questionnaire Data

(** indicates reversed thresholds)

4.2.3.1 Summary Statistics

The summary statistics for the overall person and item fit residuals are presented in Table 4-4. The residual mean value for the items is $M=-0.78$, $SD=6.55$, both of which deviate substantially from their respective expected values of 0 and 1. The misfit is further supported by the chi square item-trait interaction value ($\chi^2=428.88$ $df=72$) and a significant probability value of $p=0.00<0.01$.

Table 4-4 Overall Item and Person Fit Statistics for the Questionnaire Data

SUMMARY STATISTICS	OVERALL FIT RESIDUALS (ITEMS)	OVERALL FIT RESIDUALS (PERSONS)	TOTAL ITEM TRAIT INTERACTION	PERSON SEPARATION INDEX	% OF MISFITTING PERSONS (FIT VALUES > 2.5)
	Mean=-0.78 SD=6.55	Mean=-0.42 SD=1.28	$\chi^2=428.88$ $df=72$ $p=0.00$	0.90	1.7%

N=809, SD= Standard Deviation, χ^2 = Chi-Squared, df = degrees of freedom, p =probability

The fit residual statistics for persons are $M=-0.42$ and $SD=1.28$, which exhibit deviations from their expected values of 0 and 1, but are close enough to their expected values so as to not raise serious concerns regarding misfit. The person separation index ($PSI=0.90$) and

its traditional counterpart, Cronbach Alpha ($\alpha=0.92$), suggest that, despite its relatively short length (only 8 items), the questionnaire can reliably distinguish between persons with lower and higher background knowledge levels.

4.2.3.2 Individual Item Fit Statistics

The individual item difficulty estimates and fit statistics are presented in Table 4-5. The items are serially ordered and display a range of difficulty measures (see the location column). In line with the observations during the questionnaire scoring process, results show that item 8 (self-assessment of performance) is the most difficult item for respondents to endorse (+0.51 logits). On the other hand, item 1, which required respondents to rate their familiarity with a topic, is estimated as the easiest item (-0.54 logits).

Different fit indices are taken into account in flagging items that display misfit. Table 4-5 shows that the residual values for items 1 to 4 (highlighted in yellow) are well below the minimum criterion value of -2.5 which indicates overfit to the model. On the other hand, items 5 and 7 (highlighted in green) have very large positive residual values above the criterion level of +2.5, which indicates that the items fail to discriminate between different levels of background knowledge. Item 5 displays the largest residual value, almost twice as much as all the other items therefore requiring further examination. Broadly speaking, underfit to the model, which indicates more variation in scores than expected, is considered to be more problematic than overfit (Myford and Wolfe, 2003; Eckes, 2009).

Misfit to the model is also evidenced in chi square statistics with probability values below $p=0.05$. In order to reduce the risk of a type I error (i.e. assuming a significant difference

when there is none), the Bonferroni adjustment option was used. All items, with the exception of items 6 and 8, are highlighted as displaying significant misfit.

Table 4-5 Individual Item Fit Statistics for the Questionnaires Items

Seq	Item	Location	S.E	Residual	DF	ChiSq	DF	Prob
1	I0001	-0.54	0.05	-5.33	699.63	28.736	9	0.0007
2	I0002	-0.40	0.05	-6.99	699.63	53.082	9	0.0000
3	I0003	-0.11	0.05	-6.92	699.63	50.734	9	0.0000
4	I0004	-0.13	0.05	-5.32	699.63	35.92	9	0.0000
5	I0005	0.44	0.05	11.21	699.63	186.561	9	0.0000
6	I0006	0.43	0.05	1.853	699.63	6.421	9	0.6971
7	I0007	-0.19	0.05	4.72	699.63	39.555	9	0.0000
8	I0008	0.51	0.05	0.49	699.63	27.876	9	0.8356

Misfit of the data to the Rasch model is examined from a diagnostic perspective and in more depth in the next section in order to decide on possible remedial actions.

4.2.3.3 Diagnosing Misfit: Thresholds and Category Functioning

One possible source of misfit is disordered thresholds, or categories not functioning as intended. The threshold map for the questionnaire items in Figure 4-8 displayed reversed thresholds for items 1 and 5 (flagged with **). The threshold and category probability curves for these two items are reproduced in Figure 4-9 and Figure 4-10.

The CPCs for items 1 and 5 show that there is no region of the person location continuum in which score category 2 (*undecided*) is most likely to be endorsed. As the probability of opting for category 1 (*disagree*) decreases, it is the likelihood of endorsing category 3 (*agree*) that increases. The threshold probability curve (TPC) for item 1 shows that threshold 3, which divides the more difficult response categories of 2 (*undecided*) and 3 (*agree*), is to the left of threshold 2, which is designed to divide the easier response categories of 1 and 2. The categories are therefore not functioning in accordance to the

implicit idea that ‘the thresholds between higher level categories are more difficult than thresholds between lower level categories’ (Van Wyke and Andrich, 2006:21). On the other hand, the TPC for item 5 shows that thresholds 2 and 3 are almost superimposed on each other, suggesting that there is little or no discrimination between their respective categories.

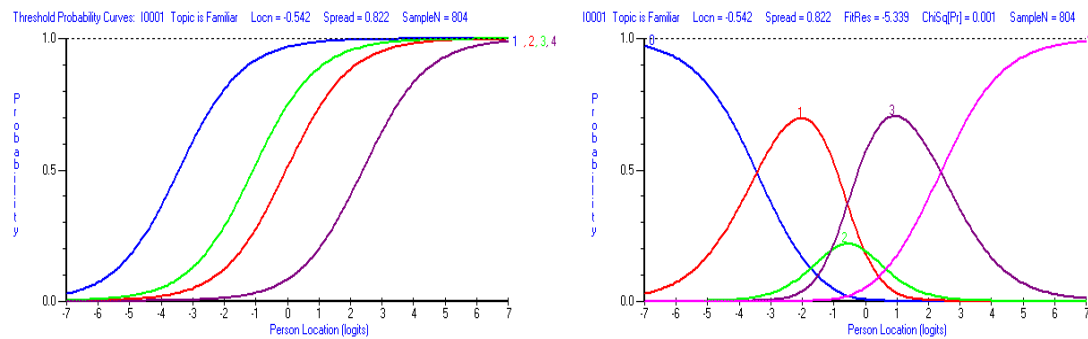


Figure 4-9 Threshold Probability Curves (left) and Category Probability Curves (right) for Item 1

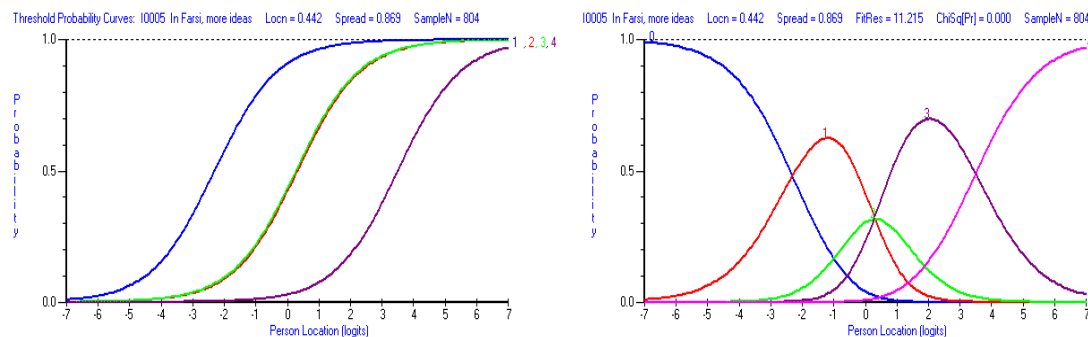


Figure 4-10 Threshold Probability Curves (left) and Category Probability Curves (right) for Item 5

The infrequency of response observations in a given category is one possible explanation for category malfunctioning. However, the category response frequencies from the RUMM output suggested that this was not the case, as there were 138 and 184 observations for the *undecided* category in items 1 and 5 respectively. An alternative explanation (Andrich, personal communication) is that the middle *undecided* category is not functioning as intended in eliciting an underlying attitude which falls between the

disagree and *agree* options on the Likert scale. The concern with designating the *undecided* or ‘not sure’ category in the middle of a response scale is voiced by Andrich, de Jong and Sheridan (1997:66-67) who, on the basis of evidence from their study of teacher attitude questionnaire data, observe that ‘the scoring of the NS [Not Sure] category as if it is operating in the middle of the other categories is not tenable’. For the current study, it has been shown empirically that the middle category is not functioning as intended for two of the questionnaire items. An approach suggested by Andrich, De Jong and Sheridan (1997:70) in dealing with disordered categories is to collapse the categories *post-hoc* and/or to consider format restructuring for further administrations of the questionnaire.

For the current study, remedial action was taken by collapsing the disordered categories. The scores for items 1 and 5 were recoded by combining the responses of score categories to form four rather than five categories: the original coding of 0,1,2,3,4 was recoded into 0,1,1,2,3. The CPCs of items 1 and 5 after collapsing of categories is provided in Figure 4-11. The categories are now ordered and the threshold map (Figure 4-12) no longer exhibits disordered thresholds. Moreover, the fit residual statistics for both items showed improvements in fit (from -5.33 to -5.30 for item 1 and from 11.21 to 8.14 for item 5) following this remedial action.

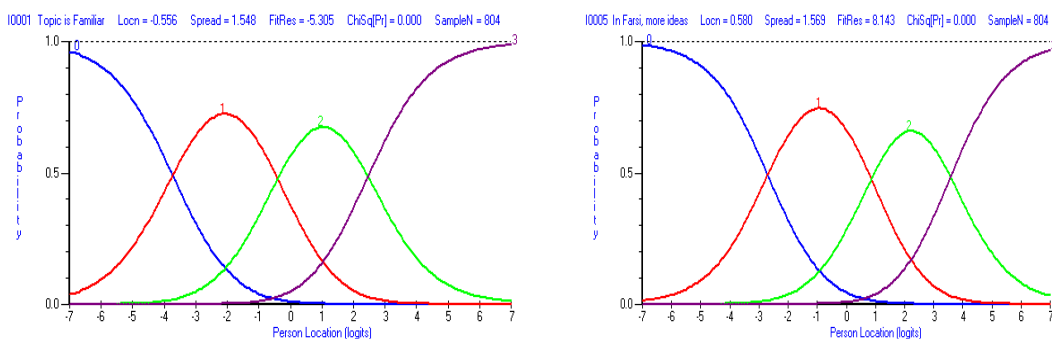


Figure 4-11 CPCs for Items 1 (left) and 5 (right) after Collapsing of Categories

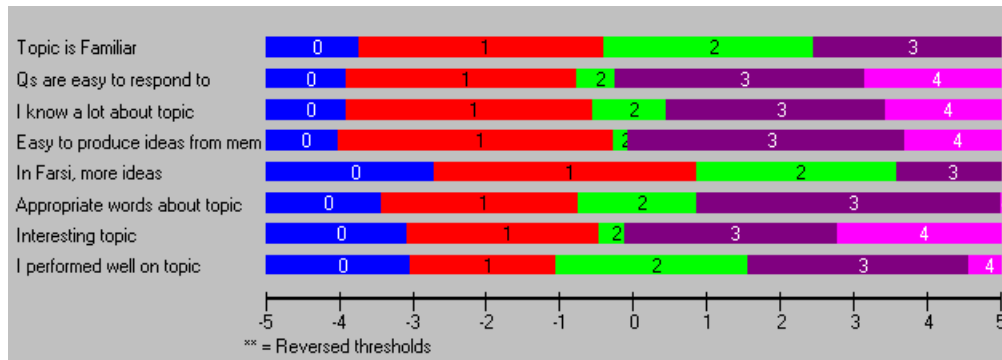


Figure 4-12 Threshold Map of for the Eight Items after Collapsing of Categories)

4.2.3.4 Diagnosing Misfit: Differential Item Functioning (DIF)

Another source of misfit that has been identified in the literature is that of differential item functioning, which occurs when ‘different groups within a sample (...) despite equal levels of the underlying characteristic being measured, respond in a different manner to an individual item’ (Pallant and Tennant, 2006:6).

On the basis of the initial observations made during the questionnaire scoring process in relation to item 5, it was hypothesised that the item might be displaying differential item functioning (DIF) for persons of different proficiency levels. In order to empirically test out this hypothesis, the person factor of *proficiency group* was first specified where C-test ability estimates were used to divide participants into four proficiency groups (low, medium-low, medium-high and high). Next, a DIF analysis was carried out for all data. Table 4-6 summarises the DIF statistics for the items exhibiting significant DIF (items 5 and 8). However, when the Bonferroni adjustment option was used, only item 5 was highlighted as displaying significant DIF.

The graphical display of item 5’s ICC, in which different proficiency levels are plotted over (Figure 4-13), clearly show the stark contrast between the way the high proficiency

persons respond to this question as opposed to the other groups. This phenomenon can be understood by re-visiting the questionnaire stem:

- *If I were to talk about this topic in Farsi, I would have more ideas to talk about*

It is likely that lower proficiency groups attribute (any) problems in performing in a second language to their language proficiency (or lack thereof), whereas for the high proficiency groups, language is no longer a barrier. Therefore, if the persons in the latter group have enough ideas, they are likely to be able to express it in their second language and if they are unable to do so, it is an indication that background knowledge-related problems might persist in their first language. This phenomenon is clearly illustrated in the ICC where the observations for the high proficiency group are well below the expectations of the model.

Table 4-6 Questionnaire Items Exhibiting Differential Item Functioning

ITEM	PROFICIENCY GROUP			
	MeanSquare	F	DF	Prob
I0005	65.21	63.56	3	0.00*
I0008	3.69	4.466	3	0.004

*=Bonferroni adjustment

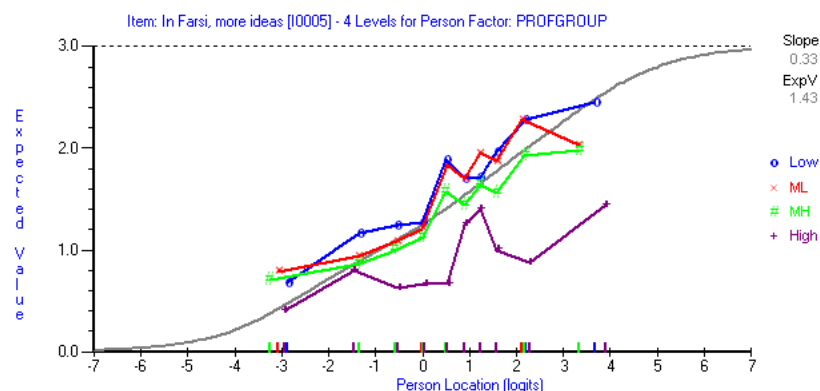


Figure 4-13 Item Characteristic Curve for Item 5 (Divided by Proficiency Level)

ML=Medium-Low, MH=Medium High

The above analysis suggests that item 5 is not working as intended for the high proficiency group. An item-split method or *item resolve* method²⁹, suggested by Andrich and Hagquist (2012) for dealing with items showing significant DIF, was adopted for item 5. However, the ICCs for the different proficiency groups after the split only showed a good fit for the low proficiency group while exhibiting inconsistencies for the remaining groups. The high proficiency group displayed significant misfit to the model with almost no discrimination along the continuum. Given the ineffectiveness of this remedial action on the one hand, and the large contribution of this item to overall misfit of data to the Rasch model on the other, a decision was made to remove the item from subsequent analyses and to re-examine item fit statistics and ICCs (see Table 4-7 and Figure 4-14).

Table 4-7 Individual Item Fit Statistics for the Questionnaire Items (After Removal of Item 5)

Seq	Item	Location	S.E	Residual	DF	ChiSq	DF	Prob
1	I0001	-0.51	0.06	-2.03	674.29	20.614	9	0.014
2	I0002	-0.40	0.05	-5.85	674.29	34.677	9	0.00*
3	I0003	-0.05	0.05	-5.40	674.29	35.054	9	0.00*
4	I0004	-0.06	0.05	-4.38	674.29	19.913	9	0.01
6	I0006	0.56	0.06	2.63	674.29	15.045	9	0.08
7	I0007	-0.16	0.05	7.44	674.29	63.434	9	0.00*
8	I0008	0.63	0.06	1.64	674.29	19.585	9	0.02

*=Bonferroni adjustment

Note: Items are serially ordered

²⁹ Resolving an item refers to creating an item specific to different levels of the person factor e.g. one item for low proficiency group, one item for medium low proficiency group and so on and so forth.

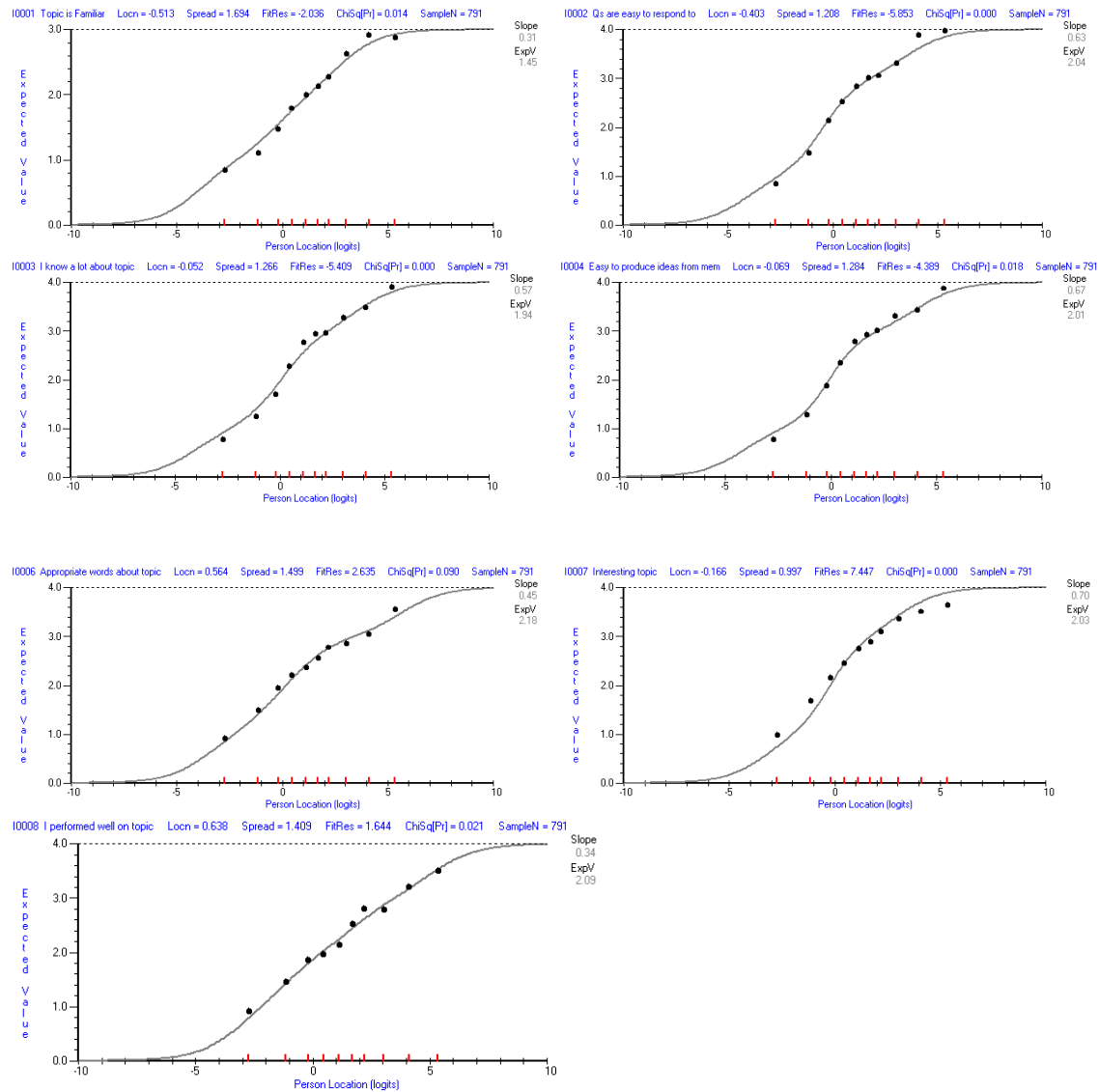


Figure 4-14 Item Characteristic Curves for Questionnaire Items 1-4 and 6-8.

The fit residual statistics in Table 4-7 still display some misfit of data to the Rasch model. However, a close examination of the ICCs for the items shows that the observations in the data are quite close to the theoretical curve and the model expectations. What should be taken into account in examining misfit is that misfit to the Rasch model is deviation from perfection and that it is the ‘practical utility’ of the model which should be taken into account (Eckes, 2009: 27):

Generally speaking, Rasch models are idealizations of empirical observations. Therefore, empirical data will never fit a given Rasch model perfectly (...). The really interesting question concerns the practical utility of a model.

The results of the analyses, the remedial actions taken, the examination of the ICCs, and the high reliability indices (PSI=0.91; Cronbach Alpha=0.92) of the background knowledge questionnaire all together suggest that the extent of the misfit of the data to the Rasch model is considered acceptable in light of the instrument’s purpose in distinguishing between higher and lower levels of topic-related background knowledge. It is therefore the interval-level background knowledge estimates from the application of the Rasch model to the questionnaire raw scores which are used for subsequent analyses in the main study.

The estimated background knowledge measures of persons on each topic (in logits) also allow for an examination of the relative familiarity of topics by considering the background knowledge means of persons for each topic; a higher mean indicates topics for which persons, in general, reported having higher background knowledge of whereas a lower topic mean indicates low levels of familiarity with a topic. Table 4-8 shows the topics, number of persons who rated their background knowledge of each individual topic, their mean (in logits) and standard deviation values. The topics have been divided by task

type and ordered serially. As it can be seen, there are differences in persons' background knowledge means on the different topics. For example, topic B.3 (Describe a Family Member) is associated with the highest background knowledge mean whereas the average background knowledge estimates on Topic C.6 (Genetic Research) is the lowest across all topics. The results of an analysis of variance (ANOVA) indicate that the difference in these means are statistically significant [$F(17,787)=14.88$, $p<0.001$]. These results suggest that the study has been successful in selecting topics for which the participants have varying levels of background knowledge.

Table 4-8 Background Knowledge Means of Persons for Each Topic

<i>Topic Ref.</i>	<i>Topic Name</i>	<i>N</i>	<i>Mean</i>	<i>SD</i>
A.1	Family	82	1.79	1.57
A.2	Leisure Time	82	2.06	1.75
A.3	Festivals	42	-0.25	1.93
A.4	Colour	42	0.97	1.42
A.5	Keeping Contact	40	1.83	1.79
A.6	Dancing	40	1.46	2.32
B.1	Describe (Friend)	42	2.09	1.61
B.2	Describe (River)	42	1.01	2.52
B.3	Describe (Family)	40	2.75	1.93
B.4	Describe (Choice)	40	1.60	2.78
C.1	Qualities of Friends	42	1.42	1.65
C.2	Other Relationships	42	-0.11	1.90
C.3	Water-based activities	42	0.24	2.09
C.4	Economic Importance of Rivers	42	-1.05	2.22
C.5	Family Similarities	40	0.80	1.79
C.6	Genetic Research	40	-1.93	2.89
C.7	Important Choices	40	1.41	2.13
C.8	Everyday choices	40	1.01	2.32

4.3 Summary

In this chapter, the psychometric properties of the study's instruments (C-tests and background knowledge of topic questionnaires) were examined using Rasch analysis. Different pieces of evidence (fit statistics, reliability indices, item characteristic curves, threshold and category probability curves, etc.) were brought together in evaluating the extent to which response data from the instruments fit the Rasch model. Remedial action was also taken, where necessary. The analyses support the raw score transformation of C-test scores and background knowledge questionnaire responses to Rasch-based interval level measures of general language proficiency and background knowledge respectively, both of which are used in the study's main analyses in the next chapter.

Chapter 5 Findings

5.1 Introduction

Chapter 5 presents the quantitative findings of this study. The aim of the chapter is to evaluate the extent to which second language spoken performance is affected by the variables of interest in the study, namely *topic* and test takers' *background knowledge of Topic* (BK). However, as discussed in the introduction and literature review chapters, the assessment of spoken performance is a complex process which can be influenced by a number of other factors or *facets* e.g. examinee abilities, tasks, interlocutors (either examiner or other candidates), rating criteria and raters as well as the interactions between them (McNamara, 1996; Eckes, 2009). For this reason, the two facets of interest in the study i.e. topic and BK need to be embedded within this larger picture in order to be better understood. To this end, the chapter starts with an introduction of a 'conceptual-psychometric framework' (Eckes, 2009:11) which is subsequently drawn upon in order to systematically report the findings from Multi-Faceted Rasch Measurement (MFRM) analysis of spoken score data. The results of other quantitative analyses (e.g. descriptive statistics of functions data, multiple regression analysis of BK and C-test data, descriptive statistics of test-taker questionnaire data) are also reported. The findings from the different analyses are discussed in more depth in relation to the research questions they pertain to.

5.2 MFRM Analysis: a Conceptual-Psychometric Framework

As discussed in the literature and methodology chapters (see sections 2.3 and 3.5.1), the classical approaches to addressing score variability concerns in rater-mediated performance assessment contexts, for example through rater training, multiple marking schemes and ascertaining high inter-rater reliability indices, were found to be inadequate. The Rasch model approach and its extension to Many-Facet Rasch Measurement (Linacre, 1989) were subsequently suggested as a sound alternative for addressing some of the limitations of the traditional approaches. To re-iterate, the strength of the MFRM approach is that it allows for all the facets of the testing situation – those held to be relevant – to be modeled, parameterised and estimated simultaneously (Andrich and Marais, 2011). The ordinal relationship between these parameters is then expressed on a common interval scale where different facets can be meaningfully compared to one another (Eckes, 2009), and examinee scores can be adjusted for the impact of other facets accordingly, thus presenting a more accurate measure of examinee ability.

In this section, a conceptual-psychometric framework by Eckes (2009:11) which has been adapted to the current study is presented in Figure 5-1. The framework serves two purposes: firstly, it provides a schematic view of the facets which are considered to be relevant to the current study and secondly, it guides the systematic reporting of data from an MFRM analysis output.

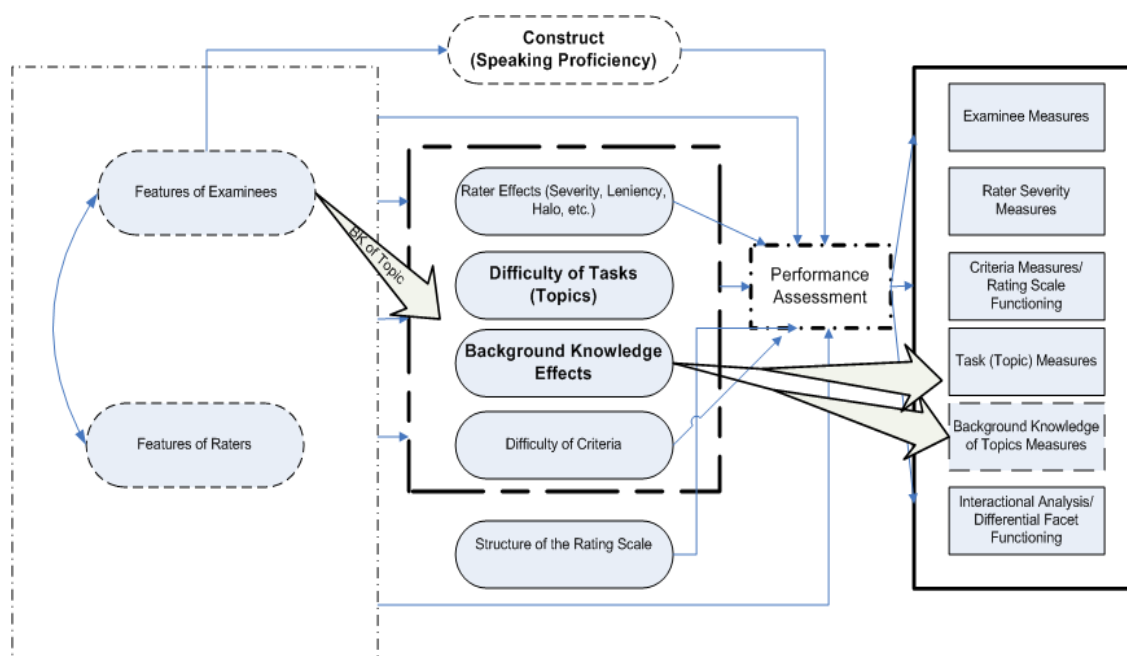


Figure 5-1 A Conceptual-Psychometric Framework of Factors Relevant to this Study (Adapted from Eckes, 2009:11)

The diagram depicts the various factors that can influence spoken performance in the current study and has been adapted from Eckes' (2009:11) original framework designed for writing performance assessment. In describing the framework, Eckes (2009:10) is careful to emphasise that 'factors shown do not encompass all that may happen in a particular rating session. The rating process is undoubtedly far more complex and dynamic that can be summarized in a diagram, and the factors coming into play are diverse at any given moment'.

The central box in the middle of the diagram represents the '*proximal*' factors which refer to those which have an 'immediate impact' on scores, the most important of which is the construct of interest, that is, the speaking proficiency of examinees, with other factors such as systematic rater effects (discussed in the literature), variability in task difficulty and rating criteria contributing to a systematic source of measurement error (Eckes,

2009:10). These '*proximal*' factors are distinguished from '*distal*' factors, identified in the box on the left as those 'exerting additional influence on the ratings, albeit usually in a more indirect and diffuse way' (*Ibid.*:10). Some of these factors have already been touched upon in previous chapters and include individual features of examinees and raters [see for example section 1.1]. It should be noted that the original framework included characteristics of the testing situation such as the technical and physical environment. However, these features were not included in the current study due to their secondary level of importance. Other factors which are normally included in a spoken performance assessment context are the characteristics of the interlocutor or examiner (O'Sullivan, 2000a; 2000b). However, in this study mechanisms were in place to control for these features to a large extent by a) having the same interlocutor administer all the tests and b) following a strict examiner script to reduce variability in test administration. For these reasons, interlocutor characteristics do not feature in the current framework as exerting substantial influence. On the other hand, background knowledge of topic, which falls under *distal* factors as a test-taker characteristic in the Eckes (2009) model, has been conceptualised in the current adapted framework as a *proximal* factor, hypothesised to exert a direct impact on spoken performance. Apart from the main influence of these two types of factors, interactions within and between them have also been conceptualised in the framework in the form of the connecting arrows.

Lastly, attention is drawn to the box on the right of the diagram which identifies the main types of measurement reports from an MFRM analysis: 'MFRM modeling generally provides detailed insight into the functioning of each factor (proximal and/or distal) that is deemed relevant in the particular assessment context' (Eckes, 2009: 11). It is the careful

consideration of various statistics in the measurement reports that allows for an understanding of the different facets as well as their interactions. Each statistic provides a different piece of information regarding the fit of the data to the Rasch model and as Tennant and Conaghan (2007:1360) aptly put, ‘each statistic gives slightly different information about the difference between the observed and expected response, similar to viewing something from a different angle’³⁰.

5.2.1 Applying the Framework: MFRM Analyses in the Current Study

In line with the adapted conceptual-psychometric framework (Eckes, 2009) discussed in the previous section, the results of three primary MFRM analyses are reported in the next sections:

- MFRM analysis with four proximal facets
- MFRM analysis with five proximal facets
- MFRM interactional or bias analyses/differential facet functioning

The findings from each set of analyses are then drawn upon to address several of the study’s research questions. Where relevant, details of preliminary or additional analyses are also presented.

5.3 Four-Facet MFRM Analysis: Results

In an MFRM analysis, the first step is to identify the different facets of the assessment situation and specify the elements within each facet. For the first analysis, the four

³⁰ It should be noted that reported statistics from different software packages are not always directly comparable (Tennant and Conaghan, 2007). Given that in the current study two different software – FACETS (Linacre, 2011b) and RUMM 2030 (Andrich et al., 2010) – were used for different purposes, information regarding the interpretation of statistics from each program will be separately provided in relevant sections.

proximal facets and the elements within each facet (as discussed in the methods chapter) are specified below:

- Examinee facet (81 participant elements)
- Rater facet (4 rater elements)
- Rating Scale (5 criteria elements)
- Speaking Tasks (18 topic elements)

In the main run of an MFRM analysis, these facets are mapped onto a common interval level scale with the computer program FACETS (Linacre, 2011b), which allows for a detailed exploration of these four different sources of variability in order to determine the extent to which these facets within this ‘complex assessment system’ are working as intended (Myford and Wolfe, 2000:1).

5.3.1 Facet Map

The vertical map in Figure 5-2 is a visual representation of MFRM where all the facets under investigation are mapped onto a single scale known as the *logit scale* (log-odd units), which is a measurement unit common to all the facets, and is arbitrarily averaged at zero. MFRM allows for all the relevant facets of a measurement situation to be ‘modeled concurrently but examined independently’ (Bond and Fox, 2007:159). This map presents, in a graphical form, the calibrations for all examinees, raters, speaking tasks (topics), rating criteria and scale categories in the study and is the most illustrative piece of output from the program.

Ability (High)		Severe	Difficult		FC	LR	GA	P	TD
Measr	+examinee	-Rater	-Topic	-Cri	S.1	S.2	S.3	S.4	S.5
4	+	+	+	+	(9)	(9)	(9)	(9)	(5)
	*							8	
3	+	+	+	+	8	8	8		
	*							---	---
2	+	+	+	+				7	
	**								4
	*				---	---	---	---	
	*								
1	+	+	+	+					
	**								---
	*								
	**			P	7	7	7	6	
	**	R2	C.6 C.4	GA					
	****	R1	C.8 C.1	FC					
:	:	R4	C.2 B.2	LR	:	:	:	:	:
:	:		C.3	:	:	:	:	:	:
*	0	*	C.5 C.7	*	*	*	*	*	3
:	:	:	A.3 A.4	:	:	:	:	:	:
:	:	:	A.1 A.5	:	:	:	:	:	:
:	:	:	A.2 B.3	:	:	:	:	:	:
:	:	:	B.1 B.4	:	:	:	---	:	:
:	:	:	A.6	:	:	:	:	:	:
	*****				---	---			
	*****	R3						---	---
	*****								---
-1	+	+	+	+					
	**			TD					
	*****				6	6	6		
	*****							5	
	*****								2
	*								
-2	+	+	+	+	---	---	---		
	*								
	*							---	
	****								---
	**								
-3	+	+	+	+	5	5	5		
	**								
	*								
	*								
-4	+	+	+	+	---	---		4	
					(4)	(4)	(4)	(3)	(1)
Measr	* = 1	-Rater	-Topic	-Cri	S.1	S.2	S.3	S.4	S.5
Ability (Low)		Lenient	Easy	Easy	FC	LR	GA	P	TD
Mean	-0.63	0.00	0.00	0.00					
S.D	1.63	0.45	0.22	0.66					

Note: Each start (*) in the second column represents 1 examinee.
FC=Fluency and Coherence, LR=Lexical Resource, GA=Grammatical Accuracy, P=Pronunciation, TD=Topic Development.

Figure 5-2 Facet Map (4-Facet MFRM)

This common scale serves as a single frame of reference for interpreting the results of the analysis:

In the **first ‘measure’ column** in Figure 5-2 we can see the logit scale or the vertical ruler onto which all the facets of measurement are located on. If data were to fit the model, then this logit scale would constitute an interval-level scale necessary for measurement (Tennant and Conaghan, 2007:1358).

The **second ‘examinee’ column** displays the examinee speaking proficiency estimates where every test taker is marked with a star (*). Notice the (+) sign next to the examinee label of the column. This denotes the ‘positive-oriented’ nature of the facet where higher measures correspond to higher raw scores for examinees (Eckes, 2009). Test takers are positioned in ascending order of speaking ability with higher ability levels appearing at the top of the column and lower ability test takers at the bottom. It is therefore possible to examine the distribution of examinees in this column. Given that the difficulty of the rating criteria is (arbitrarily) centered at zero, we can see that the examinees in this study are generally of lower abilities and are mostly clustered around 0 to -2 logits on the basis of their scores on the rating scales across criteria (columns FC, LR, GA, P and TD to the right of the map). Such a distribution is to be expected in a general speaking proficiency test where the majority of test takers fall in the middle categories of the scale.

The **third ‘rater’ column** displays the four raters in the study and their relative harshness and leniency. The (-) sign next to the ‘rater’ label of the column denotes the negative orientation of the facet where higher measures correspond to higher rater severity and lower raw scores i.e. the more severe the rater, the lower the scores he/she assigned to examinees. A glance at the rater column shows that Rater 3 is strikingly more lenient than

the other raters and is positioned towards the lenient end of the severity continuum while the other raters are located above 0 and towards the severe end of the column. Rater 2 (R2) is the most severe of the four.

The **fourth ‘topic’ column** displays the 18 speaking test tasks used in the study. Letters A, B and C denote the type of task i.e. Interview or Information Exchange, Individual Long Turn and Two-Way Discussion while the numbers denote the different topics within each task type. Similar to the ‘rater’ facet, the ‘speaking task (topic)’ facet is negatively oriented where higher measures correspond to higher task difficulty. It can therefore be seen that speaking tasks C.4 and C.8 are amongst the most difficult topics while topics B.4 and A.6 are the easiest. Topics C.7 and A.4 are estimated to be of average difficulty (close to zero).

The **fifth ‘criteria’ column**, also a negatively oriented facet, displays the analytic criteria used for scoring speaking performances. The map clearly shows that the four criteria of Fluency and Coherence (FC), Lexical Resource (LR), Grammatical Accuracy (GA) and Pronunciation (P) are of very similar difficulty levels whereas Topic Development (TD) is significantly easier, judging by its notable distance from the other criteria. In other words, TD was the easiest criterion for examinees to get a higher score on.

Lastly, the remaining five columns map the 9-category scale for the IELTS speaking criteria (FC, LR, GA and P) and the 5-category scale for the additional TD criterion to the logit scale. The lowest and highest categories for each scale are marked in parentheses, signifying extreme categories³¹. The horizontal dashed lines in these columns signify the

³¹ For the two lowest and highest extreme categories, ‘the boundaries are $-\infty$ (for the lowest one) and $+\infty$ (for the highest one)’ (Eckes, 2009:15).

category thresholds i.e. the point at which a examinee with an average expected score has a 50% probability of being assigned to one of the two adjacent categories (Andrich, 1998). Put differently, these are the points at which ‘the likelihood of getting the next higher rating begins to exceed the likelihood of getting the next lower rating’ (Myford and Wolfe, 2000:10).

At the bottom of the diagram, the mean and standard deviation of the distribution of measures (in logits) for the four facets has been reproduced. It can be seen that with the exception of the examinee facet, the mean of the remaining facets is 0.00. This is the convention in MFRM to let the examinee facet float but to centre all the remaining facets in order to ‘establish the origin of the scale’ (Myford and Wolfe, 2000:11) and to ensure that the frame of reference is ‘sufficiently constrained’ (Linacre, 1994:27).

The following sections describe the results of the MFRM for each individual facet and evaluate the fit statistics for the elements in these facets. The speaking tasks and topics are discussed last so that the research questions may be addressed within the study’s overall speaking assessment context.

5.3.2 The Examinee Facet

The examinee measurement report (N=81)³² is available in Appendix R. For illustrative purposes, the measurement results from 10 examinees displaying a range of ability levels from lowest to highest have been reproduced in Table 5-1. The summary results for the whole sample is presented at the bottom of the table.

³² The data for one participant was not available, as the digital speaking file was corrupt.

If we re-examine the facet map (Figure 5-2), we can see the wide distribution of the examinees: Table 5-1 shows that examinee 65 has the lowest speaking ability with -3.37 logits and examinee 40 has the highest speaking proficiency with +3.44 logits, spanning 6.81 logits. FACETS reports a set of *separation statistics*, provided at the bottom of the table. These are group-level statistics which are useful in summarising observations and in drawing inferences. However, it should be noted that their interpretation is dependent on the facet under examination. For examinees, these statistics can provide information regarding the extent to which the speaking tasks have been successful in separating examinees into statistically distinct speaking proficiency strata.

Table 5-1 Examinee Measurement Report (Illustrative Examples)

ID	OBSVD AVG	FAIR-M AVG	MEASU RE	MODEL S.E	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
65	3.9	3.92	-3.37	0.12	0.91	-0.8	0.89	-0.9
13	4.1	4.03	-3.1	0.15	0.99	0	0.98	0
30	4.8	4.96	-1.43	0.16	1.24	1.4	1.3	1.7
14	5.6	5.36	-0.78	0.17	1.23	1.2	1.22	1.2
56	5.8	5.89	0.17	0.19	0.89	-0.5	0.87	-0.6
17	6.1	5.91	0.21	0.17	1.07	0.4	1.06	0.3
15	6.6	6.65	1.65	0.15	1.56	3.2	1.88	4.7
67	4.9	5.04	-1.3	0.18	0.89	-0.5	0.89	-0.5
72	7.3	7.33	3.17	0.17	1.3	1.6	2.07	4.1
40	7.4	7.42	3.44	0.2	1.1	0.5	3.56	6.5
Mean (N=81)	5.4	5.41	-0.63	0.15	1.00	-.1	1.07	.0
S.D	0.7	0.75	1.36	.03	.35	1.9	.55	2.2
Model, Sample: RMSE .16 Adj (True) S.D. 1.36 Separation 8.64 Strata 11.86 Reliability .99 Model, Fixed (all same) chi-square: 6043.5 d.f.: 80 significance (probability): .00								

Note: S.E=Standard error. Infit and outfit are mean-square statistics.

Note: Observed and Fair-M Averages are on the basis of scores on all five criteria.

The first of these statistics to consider is the *separation index* (denoted by G) which is the ratio of the true standard deviation to the average measurement error (or the root mean-square error, RMSE) and it estimates the number of statistically distinguishable performance levels (Linacre, 2011a). This index is an indication of the extent to which the

test has been successful in separating examinees by their performance (Myford and Wolfe, 2000).

The separation index ($G=8.64$) suggests that there are at least 8 statistically distinct performance levels which have been identified in the sample. A closely related statistic is that of *strata* (denoted by H) which is defined as ‘the number of statistically distinguishable levels of performance in a normally distributed sample with the same “true standard deviation” as the empirical sample, when the tails of the normal distribution are modeled as extreme “true” levels of performance’ (Linacre, 2011a: 293). The choice of statistic to report depends on whether the extreme scores or outliers in the sample are ‘accidental’ or whether they represent ‘extreme performance levels’ (Wright and Masters, 2002:888). In the case of the current study, the extreme scores³³ indeed presented persons of relatively high speaking abilities or very low speaking abilities. Given the small sample size, it is not surprising that there are only a few persons at the two extremes. This is why the strata statistic provides a more accurate measure of the spread of examinee ability. This statistic is calculated as follows:

$$H = \frac{4(G) + 1}{3}$$

Results show that examinees are divided into approximately 12 statistically distinct speaking ability strata ($G=8.64$, $H=11.86$).

These separation indices are associated with a *separation reliability value* which ‘provides information about how well the elements within a (...) facet are separated in order to

³³ In the context of multiple-choice or true-false formats, extreme scores should be dealt with carefully as guessing or luck can contribute to such extreme scores. In performance-based assessment, the accidental occurrence of extreme scores is less likely.

define reliably the facet' (Eckes, 2009:20). This value gives is an indication of how different examinee measures are, and should be distinguished from *inter-rater reliability* which is an indication of how similar raters are (Myford and Wolfe, 2000). The interpretation of this index is facet-dependent: for examinees, a high separation reliability value (close to 1) is preferable³⁴ whereas a low value (close to 0) is desirable for elements in a facet which should be similar in measures e.g. for raters. The high separation reliability value of (0.99) for the examinees in the study shows that the observed differences in speaking proficiency estimates are predominantly due to differences in the underlying construct and not to measurement error (Myford and Wolfe, 2000).

The output also includes a '*homogeneity statistic*' (Eckes, 2009:19) i.e. an overall chi-square statistic and its associated probability value which test the null hypothesis that the elements in a given facet share the same measure, once measurement error has been taken into account (Linacre, 2011a). For the examinee facet, it tests the null hypothesis that examinee ability measures are the same, once measurement error has been allowed for. Results ($\chi^2=6043.5$ d.f.:80, $p=0.00<0.05$) suggest that the null hypothesis has been rejected and that variability observed in speaking ability measures is highly significant.

Another important group-level statistic is the average standard error of measurement (at the bottom of the Model S.E column), which can be interpreted in the same way as the standard error of measurement in classical test theory i.e. as an indication of the 'precision and stability of the examinee proficiency measures,' (Myford and Wolfe, 2000:14). The average standard error of measurement for the current sample is 0.15 logits. One of the strengths of the Rasch model is that unlike classical test theory where a single standard

³⁴ A good instrument should be able to separate people into different levels of ability. Items on the test should therefore have different difficulty levels to be able to elicit varying levels of the underlying ability.

error is reported for the whole sample, in MFRM a unique standard error of measurement is estimated for each examinee. The range in the current study is 0.10 to 0.21.

For illustrative purposes, the impact of the standard error of measurement on the *average* examinee can be investigated. In order to make this impact tangible in terms of the 9-band IELTS scale, the MFRM analysis was rerun without the TD criterion. The mean of the examinee speaking proficiency measure on the four IELTS criteria was -1.18 logits with an associated average standard error of measurement of 0.19 logits. We would therefore expect the true proficiency of an examinee to lie between -1.56 and -0.80 logits which correspond to approximately 5.8 and 6.2 on the IELTS scale³⁵ 95% of the time. We can therefore expect the examinees' true score to lie within four-tenth of a band score 95% of the time. However, it is emphasised that each examinee measure is associated with its individual standard error, denoting the precision with which the element measure was estimated. Generally speaking, 'the greater the number of ratings an estimate is based on, the smaller its standard error' (Eckes, 2009:15). In the examples provided in the table, the ability of examinee 65 was estimated with the highest precision (SE=0.12) while the lowest precision is observed for the estimates for examinee 40 (SE=0.20)³⁶.

Moving away from group-level statistics, we can now focus attention on individual examinees. FACETS produces a series of unstandardised fit indices which 'enable diagnosis of aberrant observations and idiosyncratic elements' (Linacre, 2011a:11)'. These

³⁵ For the calculated logit measures, the corresponding values on the IELTS scale were derived from the examinee measurement results by identifying examinees who displayed ability measures close to the calculated -1.56 and -0.80 logits. Their corresponding Fair-M average results are then reported.

³⁶ It was explained in the methodology chapter, that the 'common batch' rating design adopted in the study meant that the number of ratings per task were not equal across tasks. Those speaking tasks in the common batch would be rated by all four raters while the tasks in additional batches would be rated by just one rater which explains the differential standard errors associated with different examinees.

indices are labeled infit³⁷ and outfit mean square statistics³⁸ and have an expected value of 1 and a range from 0 to infinity where ‘the higher the (...) mean-square index, the more variability we can expect in the examinee’s rating pattern’ (Myford and Wolfe, 2000:15). Outfit mean-square is an *unweighted* mean-square for observations and is ‘particularly sensitive to occasional highly unexpected ratings’ while infit mean square is a weighted fit statistic and less sensitive to outliers or unexpected ratings but rather, more affected by ‘an accumulation of unexpected ratings’ (Eckes, 2009:17).

The standardised fit indices are a transformation of the unstandardised mean-square fit indices towards a ‘unit-normal distribution (...) effectively a t-statistic with infinite degrees of freedom i.e. a z-score’ (Linacre, 2011a:178). They can be drawn upon to infer whether ‘observed rating patterns are most likely to include surprising or unexpected ratings’ (Myford and Wolfe, 2000:18).

For the current study, it is the unstandardised information-weighted mean square statistic that is mainly reported, as it is less sensitive to outliers (compared to outfit) and because it is broadly viewed as the more important statistic to be considered in evaluating the fit of the data to the model (Eckes, 2009; Myford and Wolfe, 2003).

As mentioned earlier, the expected value for the infit statistic is 1. Values below 1 are considered to be ‘overfitting’ the model and too predictable while values above 1 are considered to be ‘misfitting’ or ‘underfitting’ the model and too unpredictable (Linacre, 2011a; Eckes, 2009). Underfit or misfit generally raises more cause for concern in comparison to overfit (Eckes, 2009). As Myford and Wolfe (2000:15) point out ‘there are

³⁷ ‘Infit’ stands for ‘information weighted fit statistic’ and ‘outfit’ stands for ‘outlier-sensitive fit statistic’ (Eckes, 2009:17).

³⁸ A mean-square fit statistic is a chi-square statistic divided by its degrees of freedom (d.f.) (Linacre, 2011a)

no hard-and-fast rules for setting upper- and lower-control limits' for the infit index. A consideration of the stakes of the test and the purposes of assessment in deciding on an acceptable range is recommended by Eckes (2009). Linacre (2011a:248) for example, suggests values between 0.5 and 1.5 as 'productive for measurement' whereas a narrower and more stringent range of 0.70 to 1.30 is suggested by other researchers (McNamara, 1996; Bond and Fox, 2007). In line with Linacre (2011a), the current study has adopted a lower control limit of 0.5 and an upper control limit of 1.5 for the infit mean square index.

The fit of the examinees to the Rasch model was therefore assessed on the basis of the examination of their infit mean square statistics. The results of examinees displaying overfit (infit index <0.5) and misfit (infit index >1.5) are reproduced in Table 5-2.

Table 5-2 Examinees Displaying Overfit or Misfit to the Rasch Model

ID	MEASURE	MODEL S.E	INFIT MNSQ	CONTROL LIMITS
49	-2.05	0.18	0.46	<0.5
5	-2.31	0.19	1.55	>1.5
15	1.65	0.15	1.56	>1.5
1	2.91	0.15	2.15	>1.5
8	2.05	0.17	2.98	>1.5

Only 5 out of 81 examinees (6%) exhibited infit indices outside of the acceptable range. The values for examinees 49 and examinees 5 and 15 are quite close to the acceptable range so as not to raise serious concerns. On the other hand, the fit indices for examinees 1 and 8 (who are amongst the top five high ability examinees) are higher than the model expectations. In order to see whether an explanation can be found for these observations, the table of unexpected responses, which is another FACETS output was examined. As its label suggests, the table produces a list of unexpected responses, identified as those

responses with standard residual values above ± 3 for any combination of facet elements. Interestingly, both examinees 1 and 8 were flagged in this table for exhibiting scores much lower than expected on certain topics for the TD criterion. This theme is discussed in more depth at a later stage.

In sum, there are only few deviations from the model expectations for a small number of examinees. It can therefore be concluded that the data from the majority of the examinees (94%) fit the Rasch model expectations to a large extent.

Returning our attention to Table 5-1, there are two column headings which have not yet been discussed: the ‘Observed’ and ‘Fair-M’ averages columns. The ‘observed average’ column displays a given examinee’s scores across tasks and raters, divided by the observed count (Linacre, 2011a). The Fair-M average, on the other hand, is the observed average, adjusted for differences in other facets such as relative rater severity or differences in task difficulty. This is a very useful statistic which is, in effect, the converted form of the Rasch measure into the raw-score metric of the original scales (Linacre, 2011a; Eckes, 2009). To illustrate the difference between these two averages, consider a comparison between two examinees’ observed raw scores: it might not be clear whether the differences are due to variability in their respective speaking proficiencies or the variability in the severity or leniency of the raters who assigned their scores or the relative difficulty of the speaking task they happened to receive. The Fair-M average addresses these problems by disentangling the influence of other facet elements from examinee speaking proficiency measures (Eckes, 2009) and thus allowing for ‘fair’ comparisons to be made.

The other advantage of the Fair-M average is the reporting of the logit measure in terms of the original scale which facilitates the interpretation of data. Given the use of the two scales in the study and in order to facilitate comparisons, the observed average, Fair-M average and standard deviations for the whole sample are separately provided for a) all five criteria, b) the 9-band IELTS scale and c) the 5-level TD scale in Table 5-3.

The table illustrates that the observed averages are very close to their adjusted Fair-M averages. In other words, the influence of the other facets on the observed averages are quite minimal. The results of the table (from the IELTS scale analysis) can also be drawn upon to compare the mean examinee ability of the Iranian participants in the study (Fair-M= 5.95, S.D=0.79) with the mean band score reported for Iran in the 2011 IELTS Test Taker Performance Report (M=6.3)³⁹, as an indication of the Iranian test taking population average for the IELTS test. It can be seen that the participants in the study display speaking proficiency scores which are, on average, only 0.35 band levels lower than reported IELTS mean band scores.

Table 5-3 Mean and S.D of Observed and Fair-M Averages for the Different Scales

ANALYSIS	OBSVD AVG		FAIR-M AVG	
	MEAN	S.D	MEAN	S.D
All Criteria	5.4	0.7	5.41	0.74
IELTS Scale	6.0	0.8	5.95	0.79
TD Scale	3.3	0.7	3.27	0.71

Note 1: The All Criteria Analysis contains all five criteria

Note 2: The IELTS Scale range is 1-9 and includes the four criteria of FC, LR, GA and P.

Note 3: The TD (Topic Development) scale range is 1-5.

In bringing the examinee results together, the analysis of different outputs of the examinee measurement report suggests that despite the study's small sample, there is a wide distribution of statistically distinct speaking ability levels with a mean average score

³⁹ Standard Deviations for the Mean Band Scores were not reported on IELTS website.

which is very close, albeit lower, than the IELTS mean band score reported for Iranian test takers. This is very much in line with the study's aims in selecting participants which are representative of the IELTS test-taking population in terms of their average ability but also the breadth of speaking proficiency. The examination of various fit statistics has shown that with a few exceptions, the majority of the examinees (94%) fit the Rasch model and that the differences observed in speaking abilities are predominantly due to variability in the underlying speaking proficiency construct and not to measurement error.

5.3.3 The Rater Facet

The literature review outlined some of the ways in which raters can introduce measurement error to performance assessment, the most problematic of which was variability in rater severity (McNamara, 1996; Eckes, 2009). Other rater-related systematic tendencies included extremism, central tendency, halo and bias effects (Myford and Wolfe, 2003; 2004; Linacre, 2011a), which were discussed in section 2.2. The next sections examine these different rater tendencies.

It should be noted that when reporting the findings, two sets of results are presented: the results of analyses with all five criteria (henceforth FullA) and the results of analyses with only the IELTS criteria i.e. without TD (henceforth IELTSA). To remind the reader, the TD criterion was included in this study as means of isolating the effects of topic on performance. By running the separate analyses, the direct linking of the study's findings to the IELTS speaking context is facilitated. Moreover, this approach allows for an exploration of the influence of this criterion in relation to the other criteria.

5.3.3.1 Rater Severity/Leniency and Consistency

In performance assessment settings, rater severity/leniency can exert a significant influence on performance (McNamara, 1996). As Myford and Wolfe (2000:1) point out, complex assessment systems require ‘going beyond interrater reliability coefficients’. As mentioned before, FACETS directly parameterises rater severity as a facet and produces severity estimates for each individual rater. The measurement reports for the raters in this study are presented in Table 5-4 (analysis of all criteria or FullA) and Table 5-5 (analysis of IELTS criteria or IELTSA) respectively and are ordered in increasing severity measures. In both analyses, Rater 3 (R3) is the most lenient rater ($R3_{FullA}=-0.65$; $R3_{IELTSA}=-1.06$) and R2 is the harshest ($R2_{FullA}=+0.34$; $R2_{IELTSA}=+0.45$), spanning 0.99 and 1.51 logits respectively. The standard error (S.E) for all rater estimates is $S.E=0.3$ in FullA and $S.E=0.4$ in IELTSA, as all raters scored a similar number of tasks and therefore their measures were estimated with the same level of precision in both analyses.

Table 5-4 Rater Measurement Report (FullA)

RATER ID	OBSVD AVG	FAIR-M AVG	MEASURE	MODEL S.E	INFIT MNSQ	OUTFIT MNSQ
R3	5.8	5.81	-0.65	0.03	0.8	1.08
R1	5.4	5.39	0.09	0.03	1.13	1.17
R4	5.3	5.31	0.22	0.03	1.12	1.17
R2	5.3	5.24	0.34	0.03	0.9	0.88
Mean (N=4)	5.4	5.44	0.00	0.03	0.99	1.08
S.D	0.2	0.26	0.45	0.00	0.16	0.14
Model, Sample: RMSE .03 Adj (True) S.D. .45 Separation 12.14 Strata 16.52 Reliability (not inter-rater) .99 Model, Fixed (all same) chi-square: 586.9 d.f.: 3 significance (probability): .00 Inter-Rater agreement opportunities: 5915 Exact agreements: 2173 = 36.7% Expected: 2187.2 = 37.0%						

Note: S.E= Standard Error, MnSq= Mean-Square R=Rater

Table 5-5 Rater Measurement Report (IELTSA)

RATER ID	OBSVD AVG	FAIR-M AVG	MEASURE	MODEL S.E	INFIT MNSQ	OUTFIT MNSQ
R3	6.4	6.52	-1.06	0.04	0.72	0.74
R1	5.8	5.87	0.30	0.04	1.12	1.17
R4	5.8	5.86	0.31	0.04	1.12	1.13
R2	5.8	5.78	0.45	0.04	0.96	0.94
Mean (N=4)	6.0	6.00	0.00	0.04	0.98	1.00
S.D	0.3	0.34	0.71	0.00	0.19	0.20
Model, Sample: RMSE .04 Adj (True) S.D. .71 Separation 18.32 Strata 24.76 Reliability (not inter-rater) 1.00 Model, Fixed (all same) chi-square: 972.1 d.f.: 3 significance (probability): .00 Inter-Rater agreement opportunities: 4732 Exact agreements: 1683 = 35.6% Expected: 1791.3 = 37.9%						

Note: S.E= Standard Error, MnSq= Mean Square R=Rater

If we look back at the facet map (Figure 5-2), we can get a general sense of the extent to which raters differ in their severity. Furthermore, the influence of rater severity on test-taker performance can also be gleaned from the map by comparing the distribution of rater severity (a spread of 0.99 logit) against the distribution of examinee ability (a spread of 6.81 logits) in FullA. In this analysis the range of examinee proficiency is 6.8 times as wide as the range of rater severity measures. When TD is dropped from the analysis, this gap is narrower, as the spread of examinee ability in the IELTSA is 8.49 which is 5.6 times the spread of rater severity (1.51). In either case, the relative distributions generally suggest that differences in rater severity are unlikely to have too large an impact on examinee scores (Myford and Wolfe, 2000). In order to further investigate the potential impact of rater severity, a useful statistic to consider is the Fair-M average for different raters. It was explained earlier that the Fair-M average is a powerful statistic which disentangles the severity of the rater from the proficiency of the examinees he/she happened to be rating. Raters can therefore be ‘fairly’ compared on the basis of Fair-M Average results. In FullA, the harshest and most lenient raters assigned scores that were

0.57 scores apart. A larger difference of 0.74 band is observed for the IELTSA. In IELTS, the smallest unit which can have a practical effect on examinee scores is half a band score (0.5 band score). As it can be seen, the difference in rater severity has exceeded this unit $0.74 > 0.50$ which means that examinees in this sample might be given scores that are half a band apart by two different raters. To further illustrate (see Table 5-5), Rater 3, on average, assigns a score of 6.52 (or 6.5 if rounded according to IELTS conversion rules) while Rater 2, on average, assigns a score of 5.78 (or 6.0 when rounded). While a difference of half a band might be considered acceptably small in operational settings, it can nevertheless have important and significant consequences for examinees, ‘particularly for those examinees whose scores lie in critical decision-making regions of the score distribution’ (Myford and Wolfe, 2000:15).

Further substantiation for differences in rater severity can be obtained from examining the group-level *separation statistics* at the bottom of the two tables. The chi-squared statistics test the null hypothesis that all raters displayed similar levels of severity when rating examinees. The chi-square values ($X^2_{\text{FullA}} = 586.9$; $X^2_{\text{IELTSA}} = 972.1$ with 3 degrees of freedom) and the associated probability values of $p=0.001$ suggest that there is 0.001 probability that the raters in the study can be thought of as interchangeable in terms of severity. The question to be posed now is whether the raters in the study have been consistent in their marking.

The rater mean square statistics can be drawn upon as indications of rater consistency: ‘rater fit refers to the extent to which a given rater is associated with unexpected ratings, summarized over examinees and criteria’ (Eckes, 2009: 16). The fit statistics in Table 5-4 and Table 5-5 demonstrate that, regardless of differences in harshness and leniency, the

raters in this study are highly consistent within themselves and they do not misfit the model. Their fit indices fall within the acceptable range of 0.5 to 1.5. In fact, even if more stringent lower and upper control limits of 0.7-1.3 are adopted, the raters still fit the model.

These findings are very much consistent with the literature on systematic rater effects, where raters are held to exhibit high levels of self-consistency in rating but nevertheless exert a significant influence on examinee scores owing to differences in severity levels (McNamara, 1996). When the results of the raters in the study, who are not trained IELTS examiners, were compared against the results of an MFRM study with trained IELTS examiners (O'Sullivan and Rignall, 2007) a smaller difference in Fair-M average scores was observed. In the current study the difference in Fair-M average scores of the harshest and most lenient examiner is 0.74 whereas this value is 1.2 in the O'Sullivan and Rignall (2007: 456) study i.e. a difference of more than one band. These results therefore suggest that with minimal training (with the *IELTS Scores Explained* DVD) and without any standardisation procedures, the raters in the study were not only consistent, but also exhibited a narrower range of severity levels than trained IELTS examiners.

The above discussion can be framed within Linacre's (2011a:13) criticism of educational testing practices where the dominant perception of inter-rater reliability is for raters to 'exactly agree with each other on the ratings provided' or, in other words, to behave like 'scoring machines'. Contrast this to assessment settings where the raters are expected to 'exhibit a specific amount of leniency or severity' and to act as 'independent experts' upon which a measurement model such as MFRM can make adjustments to raw scores for differences in severity.

A relevant output in FACETS, which can be used to interpret whether raters belong to the ‘scoring machine’ or ‘independent expert’ category, is the difference between the percentage of observed ‘exact agreements’ between raters and the Rasch model’s ‘expected agreements’. These percentages can be viewed at the bottom of the rater measurement report tables (Table 5-4 and Table 5-5). An exact agreement percentage of close to 90% is indicative of ratings awarding exact scores whereas an exact agreement percentage which is close to the model’s expected agreement is indicative of the ‘independent expert’ scenario which is conceptually, the ideal scenario in MFRM.

Results show that in both analyses, the percentage of exact agreements are very close to the Rasch model’s expectations. In FullA, of 5915 exact agreement opportunities, there is a 36.7% observed agreement which is very close to the Rasch model expectation of 37%. Similarly, of the 4732 exact agreement opportunities in the IELTSA, there is a 35.6% exact agreement which is close to the 37.9% expectation of the model. These findings suggest that the marking behaviour of the raters in the study is very much in line with the Rasch model expectations i.e. raters are acting as independent experts. While the results of the study are only based on four raters, they can still be drawn upon in making a case for using MFRM for raw score adjustments in performance based assessment contexts, particularly given the overwhelming evidence from the literature which highlights the ineffectiveness of rater training and standardisation in influencing rater harshness and leniency tendencies.

5.3.3.2 Other Systematic Rater effects: Extremism, Central Tendency and Halo⁴⁰

In addition to rater severity and consistency, other systematic rater tendencies have been identified in the literature. To remind the reader, these included *central tendency* and *extremism* referring to a rater's tendency to award scores in the middle and extreme categories of a rating scale respectively. The *halo* effect is defined as 'a rater's tendency to assign ratees similar ratings on conceptually distinct traits' (Myford and Wolfe, 2004:209) or in the words of Yorozuya and Oller, Jr. (1980:136 cited in Farrokhi and Esfandiari, 2011:1532) 'a kind of (judgment bias) spillover across scales causing them to be more strongly correlated with each other'.

Myford and Wolfe (2000:42) have suggested a set of rater effect criteria which can be used for detecting these different types of rater tendencies. Table 5-6 shows the type of rater effect and the infit and outfit limits associated with each effect in the first three columns. The remaining columns show the fit indices for each rater (separately provided for the IELTS scale and the TD scale) which can be directly compared against the defined criteria.

Table 5-6 Comparison of Rater Effect Criteria Against the Study's Rater Fit Statistics

<i>Rater Effect</i>	<i>Infit Criteria</i>	<i>Outfit Criteria</i>	<i>Rater ID</i>	<i>Infit Results (IELTS Scale)</i>	<i>Outfit Results (IELTS Scale)</i>	<i>Infit Results (TD Scale)</i>	<i>Outfit Results (TD Scale)</i>
Accurate	$0.7 \leq \text{infit} \leq 1.3$	$0.7 \leq \text{infit} \leq 1.3$	R1	1.12	1.17	1.10	1.20
Random	$\text{infit} > 1.3$	$\text{outfit} > 1.3$	R2	0.96	0.94	1.05	1.03
Halo/ Central	$\text{infit} < 0.7$	$\text{outfit} < 0.7$	R3	0.72	0.74	0.79	0.83
Extreme	$0.7 \leq \text{infit} \leq 1.3$	$\text{outfit} > 1.3$	R4	1.12	1.13	1.05	1.03

Rater Effect Criteria are based on Myford and Wolfe (2000:42)

⁴⁰ Rater interactions with other facets (another source of systematic rater effects) are not reported in this section but are covered in the section on bias analyses.

Attention is drawn to last two rows in the table. It is suggested that infit and outfit values below 0.7 (Myford and Wolfe, 2000:42) or below 0.5 (Linacre, 2011a: 249) are considered to be ‘muted’, indicating very little variation in score assignment across categories and evidence of a central tendency and/or a halo effect. In contrast, an infit index between 0.7 and 1.3 with an associated outfit larger than 1.3 are considered to be ‘noisy’ indicating unexpected and inconsistent irregularities and evidence of extreme category overuse (Linacre, 2011a; Myford and Wolfe, 2000). The results of the table suggest that the presence of these systematic rater effects is unlikely.

Marais and Andrich (2011:195) make a distinction between a halo effect – which is unique to different raters – and one which is common to all raters. They argue that the latter ‘*common rater halo effect*’ (CRHE) cannot be directly diagnosed with standard fit statistics and support their argument by drawing on evidence from a series of data simulations. In their paper they formulate the CRHE as a violation of the assumption of local independence which can be detected in contrasts between two-facet stack and rack Rasch analyses⁴¹. Andrich (2006) suggests that when different criteria share the same common rating scale structure (such as the FC, LR, GA and P criteria in the 9-band IELTS scale), a CRHE might be instigated where raters show a tendency to award profiles i.e. the assignment of same score across criteria which can in turn lead to artificial consistency and measurement error (Marais and Andrich, 2011). The CRHE was examined in the context of the current study for the analytic criteria in the IELTS rating scale using RUMM 2030 (Andrich et al., 2010). The TD criterion was not included as it does not

⁴¹ In a stack analysis ‘the same persons rated by different raters are treated as distinct persons, each identified by an identification number and a rater’. In a rack analysis, ‘the ratings of a person rated by different raters are treated as ratings to distinct criteria’ (Marais and Andrich, 2011: 195).

share a common scale structure with the remaining criteria. The methods and analyses followed the procedures outlined by Marais and Andrich (2011) and will not be repeated here due to space limitations.

Marais and Andrich (2011:209) found the CRHE to be diagnosable in the contrast in the distributions of person and threshold estimates between two-facet stack and rack analyses:

In the former [stack] the thresholds are clustered and the distribution of persons is multimodal; in the latter[rack] all thresholds are close together and the distribution of persons is unimodal. In the former, the scale is stretched and the person separation inflated, relative to the latter.

The person-item threshold distributions from the stack and rack analyses in the current study are produced in Figure 5-3 and Figure 5-4 respectively. In the stack analysis, the person-distribution (Figure 5-3) is bimodal and skewed, and the thresholds are clearly clustered in specific points. This is in line with the results of Marais and Andrich (2011:209) where ‘at the highest magnitude of dependence, the thresholds align very close to each other and the person distribution shows peaks between where the thresholds are located’. This symptom is clearly observable in the shape of the distribution in the current data. A notable difference can be observed in the shape of the distribution in the rack analysis (Figure 5-4) compared to the stack analysis, where it can be seen that the person-threshold distribution is now unimodal and the thresholds are much closer together. The scale is also far less stretched relative to the stack analysis. Additional diagnosis of the CRHE was also outlined as ‘decreased reliability and person SD when dependent ratings for each rater are summed... compared to when they are analysed separately’ (Marais and Andrich, 2011: 208). Table 5-7 shows the results of the two analyses labelled RackSumA for the analysis with summed dependent ratings and RackSepA for the analysis where

ratings were separately analysed. It can be seen that the person separation index, the person means and SDs have all decreased in RackSumA compared to RackSepA.

While the evidence presented above suggests the potential presence of a CRHE in the study, it is not possible to draw conclusive results on two grounds. Firstly, the raters in the study were briefed about the potential halo effect prior to the rating process so that they would 'attempt to guard against this tendency' as recommended by Myford and Wolfe (2003: 396). Secondly, in their post-rating interviews the raters reported a strong awareness of the potential halo effect each time they awarded flat profiles, leading them to double-check their ratings. They emphasised that when flat profiles were awarded, it was due to the observed features of the spoken performance which fit a particular band across all criteria. In other words, it was the performance and not the rating scale structure that led to similar ratings. This is a very important distinction and one which requires further research. The Marais and Andrich (2011) paper was based on simulated data and even though the findings from the experimental data in the current study match the results of the simulated data very well, it is not clear whether it is the spoken performances or a common rater halo effect (or both) which contributed to the observed results. For these reasons, no further corrective actions were implemented. Nevertheless, the potential effect of a CRHE on giving an artificial impression of consistency (Marais and Andrich, 2011) should be borne in mind.

Table 5-7 Results of the Rack Analyses

<i>Analysis Name</i>	<i>Person Separation Index</i>	<i>Person Mean</i>	<i>Standard Deviation</i>
RackSepA	0.90	-2.039	2.583
RackSumA	0.80	-0.675	1.237

(RackSepA: dependent ratings for each rater are analysed separately, RackSumA: dependent ratings for each rater are summed)

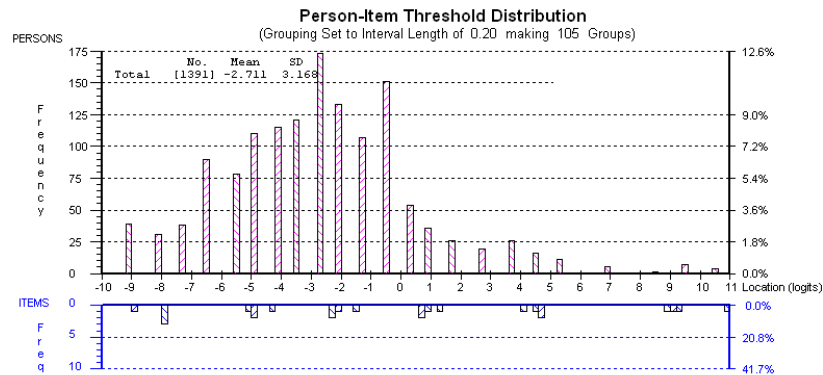


Figure 5-3 Person-Item Threshold Distribution (Stack Analysis), left

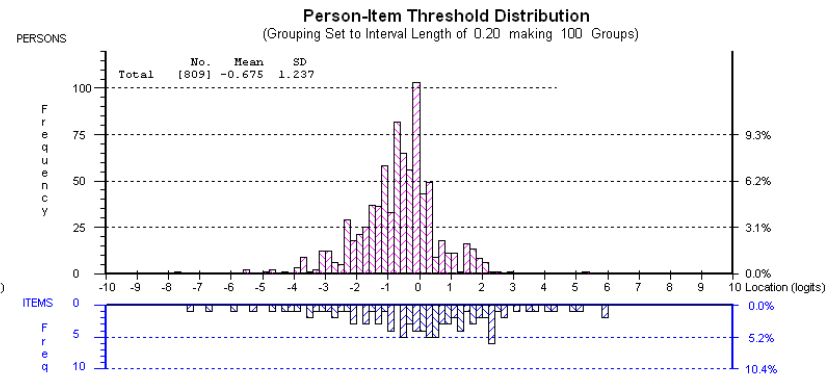


Figure 5-4 Person-Item Threshold Distribution (Rack Analysis), right

A last word on the subject and the reason why so much emphasis has been put on the CRHE is because of the concerns voiced by Marais and Andrich (2011:208), amongst others, regarding its impact on rater estimates and subsequently on person estimates when using MFRM:

Halo effect has different impacts in different analyses on the estimates. The impact of the halo effect on rater estimates was that the SD of rater estimates increased with halo effects of increasing magnitude in both the three and two-facet stack analyses. This means that, when there is a halo effect in the data, raters differed more in their severity than the simulated differences. However, in the three-facet analysis person estimates are adjusted for differences in rater severity, which is especially critical when different raters judge different persons. Therefore, if differences in severity of raters are exaggerated, person estimates will be over adjusted.

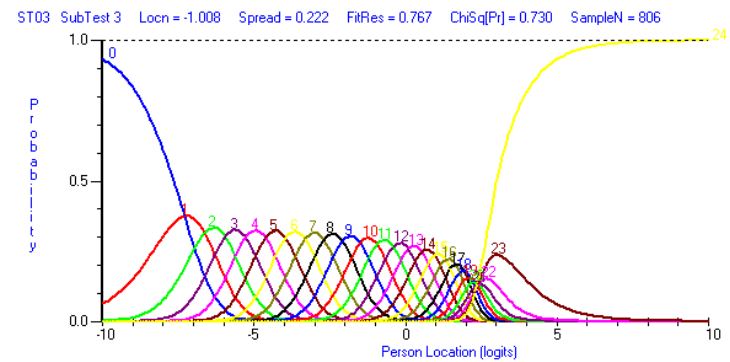
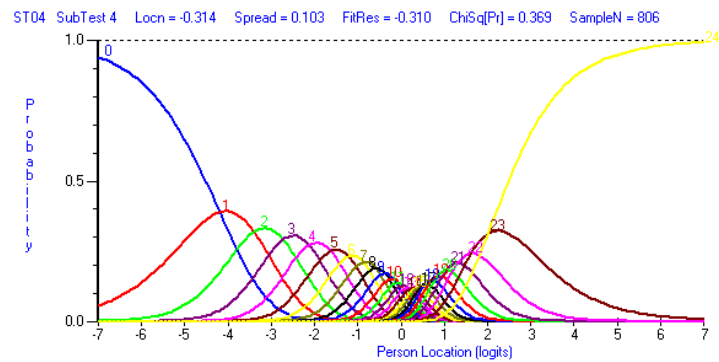
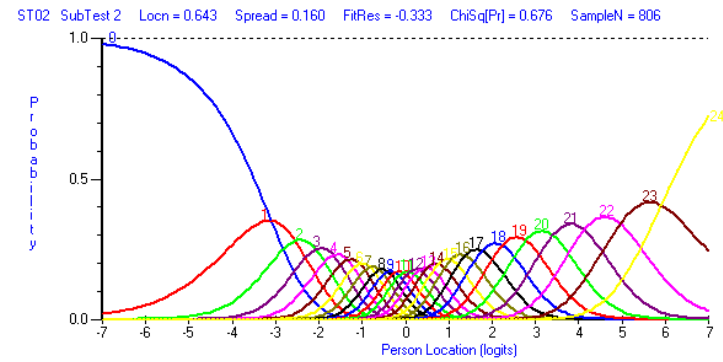
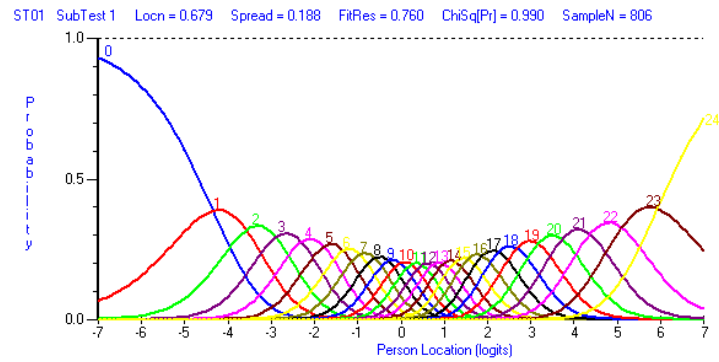
For illustration purposes and assuming that a CRHE is present in the current data, the relative rater severities from the RackSum Analysis were matched against rater severities from a three-facet Rasch Analysis (with persons, raters and criteria as the facets) and the measurement results are reproduced in Table 5-8.

Table 5-8 Rater Severity Estimates in the RackSum Analysis (left) vs. the 3-facet Analysis (right)

Rater (RackSumA)	Location	S.E	Residual	Rater (3-Facet Analysis)	Location	S.E	Residual
R2	0.67	0.03	0.76	R2	1.40	0.09	1.07
R1	0.64	0.02	-0.33	R1	1.14	0.09	-2.51
R4	-0.31	0.02	-0.31	R4	-0.96	0.08	-1.35
R3	-1.00	0.03	0.76	R3	-1.57	0.11	0.35

In both analyses, the raters are ranked in the same order in terms of severity, with Raters 2 and 1 as the most severe, Rater 4 of average severity and Rater 3 as the most lenient. However, the distances seem to have been stretched in the three-facet analysis where the difference between the harshest and the most lenient rater is 2.37 logits whereas in the case of the RackSumA, this distance is reduced to 0.99 logits. This observation reflects the concerns of Marais and Andrich (2011) discussed above.

Furthermore, an examination of the category probability curves (CPCs) for each rater's summed criteria, reproduced in Figure 5-5 can reveal how the individual raters' harshness and leniency interacts with the CRHE. For the most severe rater (R2), the CPCs are spread out across the continuum and are in an orderly fashion, displaying a slight halo effect in the middle categories. Rater 1's CPCs are compressed in the middle of the continuum suggesting that the rater is displaying the halo effect in the middle of the continuum. In other words, for low to middle ability persons, the rater has the tendency to award similar scores, which is why the probability of choosing the middle categories decreases. The CPCs for Rater 4 suggest the strongest effect of the halo, as the categories, from low to high scores, are compressed in the middle. Rater 3, on the other hand, is showing consistency in assigning scores across categories at the lower levels, with the curves beautifully lined up in the left part of the graph. However, the curves are completely compressed in the upper right end of the graph. This rater's leniency interacts with halo effect and is particularly perceptible for the higher ability groups; a person might therefore be awarded flat profiles of 4,4,4,4 or 5,5,5,5 (original IELTS band scores 7 and 8) i.e. sums of 16 and 20, which explains why the score categories in between (e.g. sums of 17,18 or 19) have much lower probabilities of being assigned. In fact, in the interviews, this rater commented that there were few high ability persons amongst the speakers and that they were 'like a breath of fresh air', which perhaps suggests that the rater was so impressed with the few higher ability speakers that flat scores at the higher band scale were automatically assigned.



Rater 2 (top left), Rater 1 (top right), Rater 4 (bottom left), Rater 3 (bottom right) [in descending order of severity]
Figure 5-5 Category Probability Curves for Raters' Summed Criteria

To summarise, the four raters in this study exercised severity levels which were significantly different from one another and cannot therefore be considered interchangeable. However, in spite of their differences and the lack of extensive training, raters exhibited high levels of self-consistency in their marking behaviour and showed no misfit to the Rasch model expectations. The lack of the classical definition of inter-rater reliability in terms of exact agreements between raters was therefore not seen as problematic as severity levels are directly parameterized in the MFRM model and person ability estimates are adjusted accordingly (Linacre, 2011a).

An examination of rater fit indices also did not reveal the influence of other systematic rater effects including extremism, central tendency and individual halo effects. However, it was suggested that a halo effect common to all raters, might exert an influence on awarded scores due to the common structure of the IETLS rating scale used in the study. Possible corrective actions for dealing with a CRHE were discussed yet not implemented in the study, as further research into the spoken performances of test takers would be required in order to ascertain that it is the CRHE, and not examinee performances, which elicited flat profiles; a research endeavor which is beyond the scope of this study. Nevertheless, this study has been successful in performing the necessary analyses for diagnosing a CRHE and showing results from real data which matched the results of studies with simulated data. While the findings are narrow in scope due to the small sample size, they can nevertheless contribute to a body of research on detecting halo effects in performance-based assessment. This would be an interesting area of further inquiry with a larger sample of raters and spoken data.

Lastly, in bringing together all the findings from the rater facet, it is concluded that the measurement error associated with systematic rater influences is to a large extent controlled for in the study and therefore the adjusted examinee scores can be considered satisfactorily reliable.

5.3.4 The Criterion Facet

Similar to the previous facets, the difficulty of each criterion has been mapped onto the linear logit scale which makes comparisons to the other facets on the same scale possible. The measurement report for the five analytic criteria used in the current study i.e. Fluency and Coherence (FC), Lexical Resource (LR), Grammatical Accuracy (GA), Pronunciation (P) and Topic Development (TD) is provided in Table 5-9 in order of increasing difficulty.

Table 5-9 Rating Scale Criteria Measurement Report

CRIT ERIA	OBSVD AVG	FAIR-M AVG	MEASU RE	MODEL S.E	INFIT MNSQ	OUTFIT MNSQ
TD	3.3	3.3	-1.15	0.04	1.43	0.89
FC	6.3	6.33	0.14	0.04	0.76	0.98
LR	6.2	6.28	0.18	0.04	0.8	1.3
GA	6.1	6.07	0.36	0.04	0.84	1.22
P	5.3	5.28	0.47	0.04	1.13	1.12
Mean (N=5)	5.4	5.45	0.00	0.04	0.99	1.08
S.D	1.3	1.28	0.66	0.00	0.29	0.46
Model, Sample RMSE .04 Adj (True) S.D. .66 Separation 18.32 Strata 24.76 Reliability 1.00 Model, Fixed (all same) chi-square: 1350.5 d.f.: 4 significance (probability): .00						

(FC=Fluency and Coherence, LR=Lexical Resource, GA=Grammatical Accuracy, P=Pronunciation, TD=Topic Development, S.D.=Standard Deviation)

The IELTS analytic criteria (FC, LR, GA, P) have similar difficulty values and are located closely to one another in the facet map (Figure 5-2), with FC as the easiest criterion ($\delta_{FC}=0.14$) and P as the most difficult criterion ($\delta_P=0.47$) i.e. only a 0.33 logit span. Markedly different, is the TD criterion ($\delta_{TD}=-1.15$) making it the easiest criterion, compared to the other criteria, for examinees to achieve a high score on. It is also the

criterion with the highest associated infit value of 1.43. While still within an acceptable quality control range of 0.5 to 1.5, its relatively higher infit value and its strikingly lower difficulty level may suggest that the criterion is functioning in a different way than the other criteria. The table shows that the infit and outfit statistics for all criteria fall within the acceptable quality control range of 0.5 to 1.5.

The *trait separation indices* at the bottom of the table ($G=18.84$; $H=24.76$) and the reliability value of 1.00 suggest almost 25 statistically distinct difficulty levels. The null hypothesis that all criteria are of similar difficulty is therefore rejected ($\chi^2=1350.$; d.f.=4, $p=0.00<0.01$). Given the marked contrast in the difficulty levels of the IELTS criteria and the TD criterion, this finding is not surprising. However, it is also important to examine the same statistics for the IELTS criteria – independent of the TD criterion – to see whether despite similar levels of difficulty, the criteria are statistically distinct in difficulty levels or not: a low separation index and a non-significant chi-square value may imply that traits are not sufficiently distinguishable (in other words redundant) or that there is a halo effect present (Myford and Wolfe, 2004).

The *trait separation indices* ($G=6.26$; $H=8.69$) and the separation reliability value of 0.98 , presented at the bottom of Table 5-10 suggest that the traits/criteria can be reliably separated into approximately 8 difficulty levels. Moreover, the significant chi-squared value ($\chi^2=118.9.$; d.f.=3, $p=0.00<0.01$) confirms that the traits do not share the same degree of difficulty and are distinguishable. These results can also be drawn upon as a further piece of counter-evidence for the presence of a halo effect (Myford and Wolfe, 2004).

Table 5-10 IELTS Criteria Measurement Report

CRIT ERIA	OBSVD AVG	FAIR-M AVG	MEASU RE	MODEL S.E	INFIT MNSQ	OUTFIT MNSQ
FC	6.3	6.37	-0.21	0.04	0.86	0.88
LR	6.2	6.31	-0.17	0.04	0.94	0.96
GA	6.1	6.10	0.07	0.04	0.95	0.97
P	5.3	5.33	0.32	0.04	1.20	1.18
Mean (N=4)	6.0	6.03	0.00	0.04	0.99	1.00
S.D	0.5	0.48	0.25	0.00	0.15	0.13
Model, Sample RMSE .04 Adj (True) S.D. .24 Separation 6.26 Strata 8.69 Reliability .98						
Model, Fixed (all same) chi-square: 118.9 d.f.: 3 significance (probability): .00						

(FC=Fluency and Coherence, LR=Lexical Resource, GA=Grammatical Accuracy, P=Pronunciation S.D.=Standard Deviation)

To summarise, the results of the criteria measurement report indicate that the different criteria fit the Rasch model and that they are statistically distinguishable in terms of difficulty and therefore not redundant. The decidedly lower difficulty level of the TD criterion compared to the other criteria and its relatively higher infit value may indicate that the criterion is measuring something conceptually different from the remaining criteria. However, given that the criterion's fit statistics still fall within the acceptable range, (any) presence of multidimensionality is not considered large enough to raise serious concerns at this stage.

5.3.5 The Topic Facet

The findings from this section are used to answer some of the study's main research questions. In order to remind the reader, a brief overview of the speaking tasks, task types and topics used in the study and in relation to the IELTS Speaking Test are repeated here from the Introduction and the Methodology chapter:

The IELTS speaking test consists of three parts, associated with three different task types: Interview or Information Exchange (Task Type A), Individual Long Turn (Task Type B) and Two-way Discussion (Task Type C). The three task types are designed to increase in

difficulty from A to C. The speaking tasks within each task type, on the other hand, are designed to be parallel in terms of difficulty. Each speaking task has a different topic which is why, in the current study, the speaking task facet is labeled the *topic* facet for ease of referencing.

In the current study, a total of 18 topics were used, distributed across task types as follows: Type A tasks: 6 topics; Type B tasks: 4 topics; Type C tasks: 8 topics (the tasks can be accessed in Figure 3-1 to Figure 3-3 in Chapter 3). The topic facet therefore has 18 elements. In the main run of the MFRM analysis, all 18 elements – regardless of task type – are included in order to get a general picture of the spread of topics and the sequencing of task types. Subsequently, additional analyses are run for each task type in order to examine the spread of topic difficulties within each task type. Similar to previous sections, separate analyses are carried out for all criteria together (FullA) and for the IELTS criteria (IELTSA). A further analysis is also run independently for the TD criterion, as it is hypothesised that topic is likely to have the largest influence on the TD scores. The results of each analysis are then discussed in relation to the relevant research questions.

5.3.5.1 Examinee Ability and Topic Difficulty Distributions

The facet map (Figure 5-2) illustrates the relatively narrow range of the topic difficulty distribution compared to the examinee ability distribution for the analysis with all the criteria. A comparison of range of examinee abilities against the range of topic difficulty measures for the three analyses with a) all five criteria b) the IELTS criteria and c) the TD criterion is summarised in Table 5-11.

The table shows that when all five criteria are included (FullA), the range of examinee ability (6.81) is 9.72 times the range of topic difficulty (0.70). This can be taken as

preliminary evidence that topic is unlikely to have a significant effect on performance. This ratio is even more pronounced once TD is removed (IELTSA), as the examinee ability range increases to 8.49 while the topic difficulty range decreases to 0.62, resulting in an examinee range which is 13.69 times the topic range. In contrast, a markedly different pattern is observed for the TD criterion analysis in which the examinee range has decreased to 5.58 while the topic difficulty range has increased to 2.12 thus reducing the ratio to 2.63.

The examinee separation indices provided in the last column of the table indicate that in all analyses, the examinees are reliably divided into statistically different ability levels, most notably in IELTSA where they are divided into 12.34 speaking proficiency strata. The separation indices are much lower for the TD criterion. Nevertheless, the criterion has discriminating power, as it can reliably distinguish between 4.7 ability strata.

Table 5-11 Examinee Range Compared to Topic Difficulty Range Across Analyses

ANALYSIS	EXAMINEE RANGE	TOPIC DIFFICULTY RANGE	EXAMINEE SEPARATION STATISTICS
All Criteria	-3.37 – +3.44 (6.81)	-0.31 – +0.39 (0.70)	G=8.64 H=11.86 Reliability .99
IELTS Criteria	-4.30 – +4.19 (8.49)	-0.37 – +0.25 (0.62)	G=9.00 H=12.34 Reliability .99
TD Criterion	-2.46 – +3.12 (5.58)	-0.97 – +1.15 (2.12)	G=3.30 H=4.74 Reliability .92

These are interesting findings which can shed light on the operation of TD in the analyses. On the one hand, the TD criterion, in itself, can reliably distinguish between different ability levels. On the other hand, its inclusion in the main analysis reduces the examinee ability range while increasing the topic difficulty range. It therefore appears that the TD criterion is functioning as intended in isolating and absorbing topic-related effects. Moreover, the results of the criteria fit statistics showed that the score observations for TD

do not deviate substantially from the expectations of the model, indicating that the TD criterion is not introducing multidimensionality to the data. These findings suggest that despite its sensitivity to the influence of topic and background knowledge of topic, TD is more strongly influenced by the underlying unidimensional speaking proficiency construct that the speaking tasks are designed to measure and is explaining various in the data. These findings lend partial support for the use of a content-oriented criterion in speaking assessment contexts (in line with Sato, 2012) and will be discussed in more detail in the final chapter.

The initial findings from the comparison of topic difficulty range in relation to examinee distribution suggest that topics are unlikely to have a large impact on scores. The topic measurement report is discussed in more detail in the next section.

5.3.5.2 Topic Measurement Report

The topic measurement reports for the FullA and the IELTSA are presented Table 5-12 and Table 5-13 respectively, where topic measures are ordered in ascending difficulty (in logits). The first column is the topic reference number, the second column identifies the task type and the third column is a short topic description while the remaining columns include similar information to those discussed for all other facets.

5.3.5.2.1 *FullA: Analysis with All Criteria*

Table 5-12 shows that when all criteria have been included, the speaking tasks used in the study have a range of difficulty measures from the easiest topic ‘Dancing’ with a logit value of -0.31 to the most difficult topic ‘Genetic Research’ with a logit value of +0.39, spanning 0.7 logits. The infit and outfit statistics for all tasks fall within the acceptable lower and upper quality control limits of 0.5 to 1.5. The separation indices provided at the

bottom of the table ($G=2.97$; $H=4.29$; Reliability= 0.90) suggest that the speaking tasks can be divided into approximately four statistically distinct difficulty strata and that there is a high degree of separation between these levels, as evidenced in the high separation reliability value ($p=0.90$). The significant chi-squared value ($\chi^2=170.3$; d.f.=17, $p=0.00<0.01$) further substantiates the high degree of separation, and rejects the null hypothesis that all topics are of equivalent levels of difficulty. These findings are not surprising, given that the task types used in the IELTS test are designed to increase in difficulty and therefore a minimum of three distinct difficulty levels (corresponding to the three task types) are to be expected. However, what needs to be further established is whether the observed progression in task type difficulty matches the expected progression: Task A measures < Task B measures < Task C measures.

An examination of the second column in Table 5-12 shows that there is not a clear and consistent pattern of progression along task types in increasing difficulty. It appears that Task Type B topics appear to be generally easier than Type A topics (with the exception of Topic B.2) which is close in difficulty to Task C topics. Some Task A topics (e.g. A.3 and A.4) exhibit difficulty measures similar to Task C topics, which is against expectations, as Task A topics are designed to be the easiest topics. Task C topics appear to closely reflect the expected pattern, as the majority of them have the highest difficulty measures. Nevertheless, the expected pattern of increasing difficulty of topics according to task type is not observed in the table.

Table 5-12 Topic Measurement Report (FullA)

TOPIC ID	TASK TYPE	DESC	OBSVD AVG	FAIR-M AVG	MEASURE	MODEL S.E	INFIT MNSQ	OUTFIT MNSQ
A.6	A	Dancing	5.9	5.62	-0.31	0.07	1.06	1.42
B.4	B	Important Choice (Describe)	5.8	5.61	-0.29	0.08	1.12	1.14
B.1	B	Friend (Describe)	5.4	5.6	-0.27	0.07	0.99	1.12
B.3	B	Family Member (Describe)	5.7	5.58	-0.23	0.07	1.03	0.99
A.1	A	Family	5.5	5.56	-0.21	0.05	1.01	1.19
A.2	A	Leisure Time	5.5	5.53	-0.15	0.05	1.06	1.22
A.5	A	Keeping in Contact	5.5	5.52	-0.13	0.07	0.86	1.14
C.5	C	Family Similarities	5.6	5.48	-0.06	0.07	0.85	0.84
C.7	C	Important Choices	5.4	5.45	-0.01	0.07	0.85	0.9
A.4	A	Colour	5.3	5.43	0.03	0.07	1.06	1.07
A.3	A	Festivals	5.3	5.41	0.06	0.07	0.92	0.91
B.2	B	River (Describe)	5.2	5.38	0.12	0.07	1.2	1.21
C.8	C	Everyday Choices	5.6	5.35	0.16	0.08	0.93	0.98
C.1	C	Qualities of Friends	5.1	5.34	0.18	0.07	0.95	0.94
C.3	C	Water based Leisure Activities	5.2	5.33	0.19	0.07	0.93	0.94
C.2	C	Other Relationships	5.1	5.32	0.22	0.07	0.92	0.92
C.4	C	Rivers-Economy	5.1	5.25	0.32	0.07	0.83	0.82
C.6	C	Genetic Research	5.4	5.21	0.39	0.07	1.21	1.47
Mean (N=18)			5.4	5.44	0.00	0.07	0.99	1.07
S.D			0.2	0.13	0.22	0.01	0.12	0.19
Model, Sample: RMSE .07 Adj (True) S.D. .21 Separation 2.97 Strata 4.29 Reliability .90 Model, Fixed (all same) chi-square: 170.3 d.f.: 17 significance (probability): .00								

5.3.5.2.2 *IELTSA: Analysis with IELTS Criteria*

Table 5-13 presents the same topic measurement report as above, but this time the TD criterion is excluded. What becomes immediately noticeable is the changes in the difficulty measures of the different topics, but also, in the way they are ranked. To illustrate, let us consider a ranking of 1 for the easiest topic and ranking of 18 for the most difficult topic. In FullA Topic A.5 was ranked 7 whereas this ranking shifted to 2 in IELTSA. Similarly, Topic B.2 ranked 12th in FullA but was subsequently ranked as the second most difficult topic (Rank 17) in IELTSA. It can therefore be seen how the inclusion/exclusion of the TD criterion can impact on both topic difficulty estimates and the ranking of tasks in terms of difficulty.

The range of topic difficulty measures is narrower in IELTSA covering 0.62 logits from -0.37 (Topic A.6) to +0.25 (Topic C.4). The fit statistics for all topics fall within the acceptable range. The separation indices provided at the bottom of the table ($G=2.21$; $H=3.27$; $p=0.83$) suggest that the speaking topics can be divided into approximately three statistically distinct difficulty strata, which is one less stratum than the results in FullA. While still relatively high, the separation reliability value has also dropped from 0.90 to 0.83 across the two analyses. Nevertheless, the null hypothesis that the different topics are of equivalent difficulty measures is rejected due to the significant chi-squared results ($\chi^2=104.9$; d.f.=17, $p=0.00<0.01$).

Table 5-13 Topic Measurement Report (IELTSA)

TOPIC ID	TASK TYPE	DESC	OBSVD AVG	FAIR-M AVG	MEASURE	MODEL S.E	INFIT MNSQ	OUTFIT MNSQ
A.6	A	Dancing	6.5	6.21	-0.37	0.09	1.01	1.11
A.5	A	Keeping in Contact	6.2	6.16	-0.27	0.09	0.85	0.91
A.1	A	Family	6	6.14	-0.23	0.06	1.03	1.07
B.1	B	Friend (Describe)	5.8	6.14	-0.23	0.08	0.97	0.96
A.2	A	Leisure Time	6	6.12	-0.19	0.06	1.15	1.16
B.4	B	Important Choice (Describe)	6.3	6.1	-0.15	0.09	1.06	1.05
B.3	B	Family Member (Describe)	6.2	6.05	-0.05	0.09	0.9	0.89
A.4	A	Colour	5.9	6.05	-0.05	0.09	1.1	1.09
A.3	A	Festivals	5.9	6.02	0	0.09	0.81	0.81
C.5	C	Family Similarities	6.2	6.02	0	0.09	0.91	0.91
C.7	C	Important Choices	6	6	0.04	0.09	0.86	0.86
C.2	C	Other Relationships	5.7	5.92	0.19	0.08	1	0.97
C.8	C	Everyday Choices	6.2	5.92	0.2	0.09	1	1.07
C.3	C	Water based Leisure Activities	5.7	5.92	0.2	0.09	0.95	0.95
C.1	C	Qualities of Friends	5.6	5.91	0.21	0.08	0.94	0.93
C.6	C	Genetic Research	6	5.9	0.23	0.09	1.1	1.12
B.2	B	River (Describe)	5.7	5.9	0.23	0.09	1.08	1.09
C.4	C	Rivers-Economy	5.7	5.89	0.25	0.09	0.82	0.82
Mean (N=18)			6.0	6.02	0.00	0.08	0.97	0.99
S.D			0.2	0.10	0.20	0.01	0.10	0.11
Model, Sample: RMSE .08 Adj (True) S.D. .19 Separation 2.21 Strata 3.27 Reliability .83 Model, Fixed (all same) chi-square: 104.9 d.f.: 17 significance (probability): .00								

When examining the progression in difficulty levels of topics in terms of task type, Table 5-13 indicates a more consistent trend compared to the previous analysis, in that with the exception of Topic B.2, Task Type C topics exhibit the highest difficulty levles. Moreover, Task Type B topics are no longer clustering at the top of the table (the location for the easiest topics). Rather, their locations are shifted down while still interspersed with

Task A topics. Once again and against expectations, tasks A.3 and A.4, which are designed to be the easiest task types (A), are closest in difficulty to Task C types.

The above findings can be drawn upon to answer the following research questions:

- i. *To what extent are the topics (speaking tasks) used in parallel versions of a language proficiency interview similar in terms of difficulty?*
- ii. *To what extent does the observed progression of topic difficulty measures match the intended progression of topic difficulty from easy to difficult?*

The topic measurement reports from the two analyses (FullA and IELTSA) demonstrate that the different topics used in the IELTS Speaking Tests can be reliably divided into approximately 3-4 statistically distinct difficulty levels. This is as expected in light of the fact that the tests are designed to include three task types of increasing difficulty. However, the sequencing of topics in difficulty measures did not match the expected progression in task type difficulty.

The most consistent pattern was observed for Task Type C (expected to be the most difficult task type) where the majority of its associated topics also exhibited the highest difficulty levels. The distinctions between Task A and Task B topics, on the other hand, were not clear-cut and their respective topics did not cluster around particular difficulty measures. In other words, the intended sequencing of topics according to task types, that is, from information exchange to two-way discussion and from familiar to abstract was not consistently observed in the data. This is a finding which may call the cognitive validity of the speaking tasks into question, as the tasks do not necessarily increase in difficulty in the same way that the test developers had intended. However, it should be noted that the

trends observed in IELTSA were more in line with the expected patterns compared to FullA, once again indicating that the TD criterion exerts an influence on the sequencing of topics within task types.

5.3.6 Rating Scales

So far, the rating scales used in the study have not been examined. An important quality consideration is the extent to which the rating scales are functioning as intended and there are various statistics that can be used in this respect.

First, let us look at the structure of the rating scales for each criterion. A visual display is available in the facet map (Figure 5-2). For this study, the Partial Credit Model was adopted in the MFRM analysis which assumes that each criterion has a unique category structure. It is distinguished from the Rating Scale model which assumes that the same category structure is common to all criteria (Myford and Wolfe, 2000). An examination of the facet map reveals the IELTS criteria to have a more or less similar category structure, although the P criterion exhibits a slightly different structure compared to FC, LR and GA. As expected, the category structure for TD is different from the other IELTS criteria because of its fewer number of score categories.

FACETS produces a category statistics output for each criterion. Due to space limitations, the output for the first criterion (FC) is presented and discussed in detail in order to highlight key aspects that need to be considered in examining similar outputs for the remaining criteria.

The category statistics for the Fluency and Coherence criterion is provided in Table 5-14. The first column shows the IELTS bands (i.e. those bands observed in the data) followed

by the count and percentage of observations for each score. It should be noted that low frequencies of observations for a particular category might contribute to misfit. In this study, the two extreme levels of the scale (Bands 4 and 9) are expected to have low frequencies.

Table 5-14 Category Statistics: Fluency and Coherence

DATA			QUALITY CONTROL			RASCH-ANDRICH THRESHOLDS	
Score	Used	Counts (%)	Avg Measure	Exp Measure	Outfit MnSq	Measure	S.E.
4	83	6%	-2.92	-2.68	0.7		
5	257	18%	-1.98	-1.86	0.7	-3.4	0.13
6	446	32%	-1.08	-1.07	0.7	-2.01	0.08
7	438	31%	-0.19	-0.24	0.8	-0.66	0.07
8	165	12%	1.48	1.25	0.8	1.4	0.1
9	13	1%	2.73	2.84	1	4.67	0.3

A key indicator of rating scale effectiveness is the average measure column, which displays the average examinee ability measures in each score category. The expectation is for these average measures to ‘advance monotonically with categories; that is, the higher the category, the larger the average measure. When this requirement is met, it is safe to conclude that higher ratings correspond to ‘more’ of the variable being measured’ (Eckes, 2009:26). The Rasch model expectations for average measures at each score category (if data were to fit the model) are also provided. Ideally, these average and expected values should be close to one another. However, if the differences were large, this misfit would be captured in the outfit mean square statistic column which, as a rule of thumb, should not exceed 2.00 (Eckes, 2009).

A final consideration is the Rasch-Andrich thresholds (or step calibrations as used in FACETS). As explained earlier, these refer to ‘the point on the examinee proficiency scale at which the probability curves for adjacent categories intersect’ (Myford and Wolfe,

2000:23). Similar to the average measures, the expectation is for these thresholds to also advance monotonically (Eckes, 2009). When categories are functioning as intended ‘the thresholds and the categories they define are naturally ordered in the sense that the threshold defining the two higher categories of achievement is of a greater difficulty than the threshold defining the two lower categories of achievement’ (Van Wyke and Andrich, 2006:14).

The statistics in Table 5-14 show that the FC scale is functioning as intended, as the observed average measures increase monotonically, exhibit values close to the Rasch model expected values and have outfit values below 1.5. There is also no evidence of disordered thresholds. The rating scale functionality can also be visually examined (see Figure 5-6) where the horizontal axis represents examinee ability while vertical axis provides the probability of category observations. Each category is represented by a curve (with different colours). As it can be seen, as the probability of a score of 4 (red curve) decreases, the probability of a score of 5 (blue curve) increases. The category thresholds are the points at which curves cross and we can see that they are ordered from left to right (Eckes, 2009).

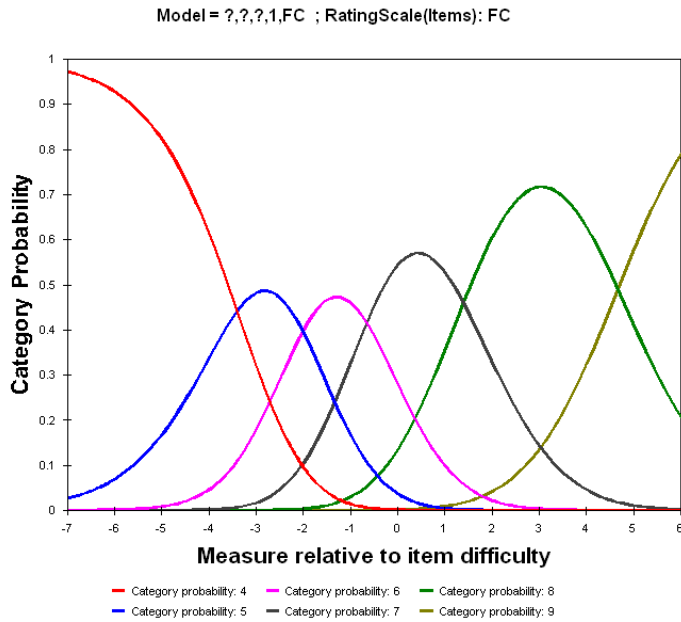


Figure 5-6 Category Probability Curves (Fluency and Coherence)

The category statistics and probability curves for the remaining criteria from the main MFRM analysis are provided in Table 5-15 to Table 5-18 and in Figure 5-7.

The category statistics and category probability curves generally show that the average measures and the Rasch-Andrich thresholds for the criteria increase in magnitude as rating scale categories increase in scores. The majority of the outfit statistics also do not exceed 2.0. The category probability curves also display ordered categories. All of this suggests that for these criteria the rating scale is functioning appropriately.

Table 5-15 Category Statistics: Lexical Resource

DATA			QUALITY CONTROL			RASCH-ANDRICH THRESHOLDS	
Score	Used	Counts (%)	Avg Measure	Exp Measure	Outfit MnSq	Measure	S.E.
4	83	6%	-2.97	-2.7	0.8		
5	271	19%	-1.98	-1.87	0.8	-3.48	0.13
6	465	33%	-1.07	-1.08	0.8	-2.01	0.08
7	403	29%	-0.17	-0.25	0.7	-0.54	0.07
8	168	12%	1.33	1.23	1	1.28	0.1
9	12	1% 1	2.62	2.81	1.1	4.74	0.31

Table 5-16 Category Statistics: Grammatical Accuracy

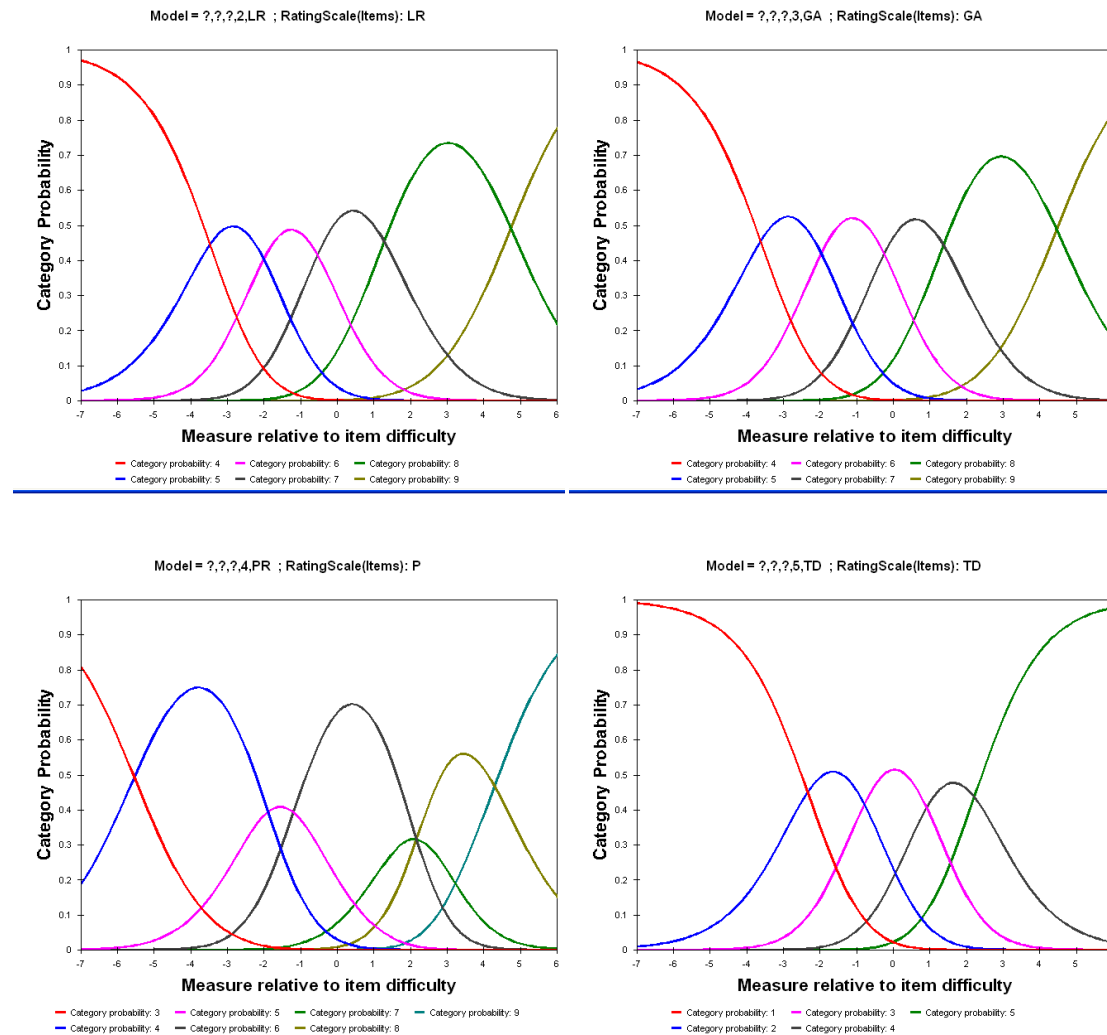
DATA			QUALITY CONTROL			RASCH-ANDRICH THRESHOLDS	
Score	Used	Counts (%)	Avg Measure	Exp Measure	Outfit MnSq	Measure	S.E.
4	94	7%	-3.08	-2.8	0.8		
5	329	23%	-2.01	-1.93	0.8	-3.62	0.12
6	513	37%	-1.08	-1.12	0.8	-1.96	0.07
7	324	23%	-0.18	-0.24	0.8	-0.25	0.07
8	130	9%	1.37	1.3	1	1.37	0.12
9	12	1% 1	2.46	2.7	1.1	4.46	0.31

Table 5-17 Category Statistics: Pronunciation

DATA			QUALITY CONTROL			RASCH-ANDRICH THRESHOLDS	
Score	Used	Counts (%)	Avg Measure	Exp Measure	Outfit MnSq	Measure	S.E.
3	21	1%	-2.85	-3.12	1.1		
4	363	26%	-2.18	-2.24	1.2	-5.54	0.23
5	397	28%	-1.34	-1.38	1	-1.89	0.07
6	505	36%	-0.6	-0.5	1.2	-1.2	0.07
7	67	5%	1.22	0.91	0.9	2.17	0.13
8	41	3%	2.21	2.29	1	2.16	0.2
9	8	1%	2.08*	2.95	1.9	4.3	0.39

Table 5-18 Category Statistics: Topic Development

DATA			QUALITY CONTROL			RASCH-ANDRICH THRESHOLDS	
Score	Used	Counts (%)	Avg Measure	Exp Measure	Outfit MnSq	Measure	S.E.
1	70	5%	-1.45	-1.42	1.2		
2	264	19%	-0.46	-0.58	2	-2.34	0.14
3	491	35%	0.51	0.22	2.9	-0.79	0.08
4	363	26%	0.95	1.04	1.4	0.92	0.07
5	214	15%	1.88	2.51	1.8	2.21	0.1



Lexical Resource (Top Left); Grammatical Accuracy (Top Right); Pronunciation (Bottom Left); Topic Development (Bottom Right)

Figure 5-7 Category Probability Curves (LR, GA, P, TD)

The main exception appears to be for the Pronunciation (P) criterion, where a problem can be immediately detected in the disordered categories, and is further evidenced in the disordered thresholds in the Rasch-Andrich thresholds column for pronunciation (marked with *). It appears that there is no region in the ability continuum in which a Band 7 is

most likely to be observed. An even closer examination reveals that the category peak for a score of 5 is also much lower than adjacent scores. One explanation for this lack of category functioning can be explained by referring back to the IELTS rating scale. The pronunciation scale is the only scale where the odd numbered bands (3,5,7) do not have clear descriptions and instead refer to the adjacent band score descriptions. Take the band 5 description (P): ‘shows all the positive features of band 6 and some, but not all, of the positive features in band 8’. It is clear from the graphical display that the raters were more likely to opt for scores where the descriptions are clear and more specific. This observation was also substantiated in the post-rating interviews with the raters, in which all 4 raters described the pronunciation criterion as the hardest to rate and referred to it as the ‘most vague’ criterion⁴².

The only other concerning statistics, are the TD outfit statistics of scores 2 and 3 (Table 5-18) which exceed the upper control value of 2.00 suggesting that there is more unexplained variability or ‘noise’ (Linacre, 2011a: 249) in the ratings.

Apart from using the category statistics in examining scale functionality, the same information can be used for a different purpose, that is, in examining the effects of topic difficulty on scores. This can be achieved by considering differences in topic difficulty levels *in relation* to average person abilities at particular score categories. This is a more fine-grained approach which allows for examining the topic effect not only on individual criteria but at the category level across criteria. This approach is first used to examine

⁴² One solution for dealing with the disordered thresholds is to collapse the categories. However, as pointed out, the problem is rooted in the original scale and not necessarily the ratings of raters. This is why the original Pronunciation scale should be further examined and perhaps modified.

topic effects at the *test level* before applying the same analyses at the *task level*. Once again, the results are separately reported for FullA and IELTSA.

5.3.7 Topic Influence on Performance Scores (Test Level)

5.3.7.1 FullA: Analysis with All Criteria

In section 5.3.6, the average examinee measures at different score categories for the various criteria were provided. The topic measurement report for the FullA was also provided in Table 5-12. Before proceeding to the analysis, the reader is reminded that in the operational IELTS test, a mark is awarded at the end of the test and for all the speaking tasks covered in the three parts of the test. In contrast, in the current study marks were awarded for each speaking task so that topic measures could be estimated separately. Given that each form of the IELTS Speaking Test is comprised of two Task A topics, one Task B topic and two Task C topics, then a combination of topics (from each task type) can be used to create multiple parallel forms. However, for illustration purposes, a decision was made to construct two parallel forms: one containing the easiest topics (within each task type) and one containing the most difficult topics (within each task type). The average differences between the two forms would then represent the maximum difference, the influence of which could be examined in relation to the average measures at different score categories across criteria.

Table 5-19 demonstrates how the two *Easy* vs. *Difficult* parallel forms have been constructed from the combination of the correct number of topics within each task type⁴³. The average difficulties for the Easy and Difficult forms are -0.17 logits and 0.18 logits

⁴³ In the operational IELTS test, the Task C topics are thematically linked to the Task Type B topic. The constructed parallel forms in this example are for illustration purposes only which is why one of Task C topics in the 'difficult form' are not the one thematically linked to the task type B topic.

respectively bringing the difference between them to 0.36 logits; a difference attributed to selected topics.

If we now consider the differences in average examinee ability measures at adjacent score categories across criteria, we can infer whether this 0.36 logit difference can have meaningful (practical) implications in terms of the rating scales. To illustrate, if we look back at the category statistics for the Fluency and Coherence (Table 5-14) we see that the average measure at Band 4 is -2.92 and the average measure at Band 5 is -1.98, implying that an increase in ability of 0.94 logits is necessary to move from Band 4 to Band 5. The difference of 0.36 logits (contributed to topic difficulty across two easy and difficult forms) is not sufficient for moving across these two adjacent bands.

The increase in average ability measures for all band scores across criteria was calculated and is summarised in Table 5-20. It is now possible to look at differences in topic difficulty levels in relation to average person abilities at particular score categories at the form level.

Table 5-19 Easy vs. Difficult Parallel Forms of the IELTS Speaking Test (FullA)

PARALLEL FORM (EASY)					PARALLEL FORM (DIFFICULT)				
Topic ID	Task Type	Desc	Fair-M Avg	Measure	Topic ID	Task Type	Desc	Fair-M Avg	Measure
A.6	A	Dancing	5.62	-0.31	A.4	A	Colour	5.43	0.03
A.1	A	Family	5.56	-0.21	A.3	A	Festivals	5.41	0.06
B.4	B	Important Choice (Describe)	5.61	-0.29	B.2	B	River (Describe)	5.38	0.12
C.5	C	Family Similarities	5.48	-0.06	C.4	C	Rivers-Economy	5.25	0.32
C.7	C	Important Choices	5.45	-0.01	C.6	C	Genetic Research	5.21	0.39
Mean (N=5)	Easy Test Form		5.54	-0.17	Mean (N=5)	Difficult Test Form		5.33	0.18
Difference in Average Measures of the Two Forms = 0.18- (-0.17)= 0.36 (logits)									

Table 5-20 Increase in Average Ability Measures in Adjacent Categories for All Criteria

BANDS	FC	LR	GA	P⁴⁴	TD (LEVELS)	
Bands 3-4				0.67		
Bands 4-5	0.94	0.99	1.07	0.84		
Bands 5-6	0.9	0.91	0.93	0.74		
Bands 6-7	0.89	0.9	0.9	1.82	Levels 1-2	0.99
Bands 7-8	1.67	1.5	1.55	0.99	Levels 2-3	0.97
Bands 8-9	1.25	1.29	1.09	0.74	Levels 3-4	0.44
					Levels 4-5	0.93
Average	1.13	1.12	1.11	1.03	0.83	

A close examination of Table 5-20 shows that there are no adjacent band levels where the average ability measure required for moving across adjacent bands exceeds the logit value of 0.36 (attributed to differences in topics across parallel forms). For example, in the FC column, the smallest average ability necessary for moving from Bands 6 to 7 (0.89) exceeds the maximum difference of 0.36 between the two test forms ($0.89 > 0.36$). Even for the easiest criterion (TD), the smallest average ability measure required to move across two adjacent bands is observed for levels 3-4 (0.44) which, once again, exceeds 0.36 ($0.44 > 0.36$). These results suggest that even when the two speaking test forms include the easiest versus the most difficult topics, the performance measures are unlikely to be influenced by differences in topic difficulties.

5.3.7.2 IELTSA: Analysis with IELTS Criteria

The same analysis is now repeated for the IELTS criteria (IELTSA). Once again, two test forms consisting of the easiest and the most difficult topic combinations are constructed, their differences calculated and compared against average ability measures necessary to move along adjacent band score categories across the IELTS criteria. In this analysis, the

⁴⁴ Due to the problems with reversed thresholds for the Pronunciation criterion, the observed average value was replaced with the expected value for the disordered categories.

Fair-M average results can also be drawn upon in providing more tangible evidence in terms of the original IELTS scale.

The difference in average topic measures in the easiest vs. the most difficult speaking test forms is 0.31 logits (see Table 5-21). This difference can also be understood in reference to the original IELTS scale where the mean Fair-M average across the easiest topics is 6.11 (rounded to 6.0 according to IELTS conventions) while the mean Fair-M average across the most difficult topics is (5.95, also rounded to 6.0). The band-level difference between these two constructed forms is approximately 0.1 IELTS scale band and far from exceeding the smallest unit (0.5 band) which can have a meaningful impact on performance scores.

Table 5-21 Easy vs. Difficult Parallel Forms of the IELTS Speaking Test (IELTSA)

PARALLEL FORM (EASY)					PARALLEL FORM (DIFFICULT)				
Topic ID	Task Type	Desc	Fair-M Avg	Measure	Topic ID	Task Type	Desc	Fair-M Avg	Measure
A.6	A	Dancing	6.21	-0.37	A.4	A	Colour	6.05	-0.05
A.5	A	Keeping in Contact	6.16	-0.27	A.3	A	Festivals	6.02	0.00
B.1	B	Friend (Describe)	6.14	-0.23	B.2	B	River (Describe)	5.9	0.23
C.5	C	Family Similarities	6.02	0.00	C.6	C	Genetic Research	5.9	0.23
C.7	C	Important Choices	6.02	0.00	C.4	C	Rivers-Economy	5.89	0.25
Mean (N=5)	Easy Test Form		6.11	-0.17	Mean (N=5)	Difficult Test Form		5.95	0.13
Difference in Average Measures of the Two Forms =0.13 – (-0.17) = 0.31 (logits)									
Differences in Fair-M Average Measures = 6.11-5.95 = 0.16 IELTS band score									

If we now turn our attention to the information in Table 5-22 we can immediately see that, on average, an increase in ability of 1.42 logits is required to move across adjacent band scores in the different IELTS criteria. The 0.31 logits of difference in topic difficulty measures is therefore too small compared to the underlying ability required for test takers

to move to a higher band. Put differently, test takers' performance measures are unlikely to be influenced by the difficulty level of topics.

Table 5-22 Increase in Average Ability Measures in Adjacent Categories for the IELTS Criteria

BANDS	FC	LR	GA	P⁴⁵
Bands 3-4				0.82
Bands 4-5	1.04	1.08	1.16	1.1
Bands 5-6	1.1	1.11	1.19	1.03
Bands 6-7	1.13	1.17	1.19	2.39
Bands 7-8	2.23	1.94	2.02	1.34
Bands 8-9	1.72	1.81	1.54	1.18
Average	1.44	1.42	1.42	1.41

Another striking observation is the differences in 'average' ability measures required to move across adjacent bands in the FullA vs. the IELTSA where, in the former, the logit measure hovers around 1 whereas in the latter, the logit measure hovers around 1.4. This finding implies that when the TD criterion is included, it requires less of the underlying ability to move across band levels, that is, it has a facilitating effect on achieving higher scores.

The findings from the above analyses can be drawn upon to answer the following research question:

iii. To what extent are parallel forms of a language proficiency interview (consisting of different topics) comparable in terms of difficulty? Are (any) differences in form difficulty large enough to have practical significance in terms of test performance?

The above analyses have demonstrated that there are very small differences in the difficulty of parallel forms, attributable to differences in topic difficulty measures. These differences are unlikely to have a significant practical influence on performance at the

⁴⁵ Due to the problems with reversed thresholds for the Pronunciation criterion, the observed average value was replaced with the expected value for the disordered categories.

criterion level or at the adjacent band score levels within criteria. If the combination of topics from two extreme difficulty levels cannot affect a drop or increase in scores at the band level, it can be inferred that, in general, differences in topic difficulties within the assessment context under study are unlikely to have a significant and practical influence on performance at the test level.

5.3.8 Topic Influence on Performance Scores (Task Level)

The sections focuses on topic measures within each task type in order to answer the following research question:

- iv. When task type is held constant, to what extent are the different topics used in parallel versions of a task similar in terms of difficulty? Are (any) differences in topic difficulty measures large enough to have practical significance in terms of test performance?*

The first part of the research question can be answered by examining the topic measurement reports for each task type. The second part of the question can be answered by repeating the same analytical procedures used in the previous section (for topics at the form level), but this time investigating the differences between the easiest and most difficult topics at the task level.

The analyses were carried out with and without the TD criterion. Only the results pertaining to the topic facet are reported in the following sections.

5.3.8.1 Task Type A: Topic Measurement Reports

The topic measurement reports for Task Type A topics are reproduced in Table 5-23 and Table 5-24 for FullA and IELTSA respectively. All topic infit and outfit statistics fall within the stringent range of 0.7 to 1.3.

The topic column in both tables shows the same ranking of topics in ascending order of difficulty (in logits), with topic A.6 (Dancing) as the easiest topic and topic A.3 (Festivals) as the most difficult topic. However, the spread of difficulty measures is wider in FullA spanning 0.3 logits whereas this range is limited to 0.10 logits in IELTSA. In determining whether the observed range in topic measures is associated with statistically distinct difficulty levels, the separation statistics at the bottom of the two tables were examined.

Table 5-23 Task Type A: Topic Measurement Report (FullA)

TOPIC	OBSVD AVG	FAIR-M AVG	MEASURE	MODEL S.E	INFIT MNSQ	OUTFIT MNSQ
A.6	5.9	5.59	-0.14	0.08	1.01	1.11
A.1	5.5	5.57	-0.12	0.05	1.01	1.08
A.2	5.5	5.55	-0.07	0.05	1.01	1.07
A.5	5.5	5.48	0.05	0.07	0.85	0.96
A.4	5.3	5.44	0.13	0.07	0.95	0.96
A.3	5.3	5.42	0.16	0.07	1.1	1.11
Mean (N=6)	5.5	5.51	0.00	0.07	0.99	1.05
S.D	0.2	0.07	0.13	0.01	0.08	0.07
Model, Sample: RMSE .07 Adj (True) S.D. .11 Separation 1.63 Strata 2.51 Reliability .73 Model, Fixed (all same) chi-square: 17.8 d.f.: 5 significance (probability): .00						

Table 5-24 Task Type A: Topic Measurement Report (IELTSA)

TOPIC	OBSVD AVG	FAIR-M AVG	MEASURE	MODEL S.E	INFIT MNSQ	OUTFIT MNSQ
A.6	6.5	6.14	-0.10	0.09	0.96	1.00
A.1	6	6.13	-0.09	0.06	1.00	1.03
A.2	6	6.12	-0.06	0.06	1.09	1.09
A.5	6.2	6.09	0.00	0.09	.84	0.88
A.4	5.9	6.04	0.1	0.09	.81	0.81
A.3	5.9	6.02	0.14	0.09	1.15	1.14
Mean (N=6)	6.1	6.09	0.00	0.08	0.97	0.99
S.D	0.2	0.05	0.1	0.01	0.13	0.13
Model, Sample: RMSE .08 Adj (True) S.D. .06 Separation .76 Strata 1.34 Reliability .36 Model, Fixed (all same) chi-square: 7.7 d.f.: 5 significance (probability): .17						

In FullA, the topic separation indices at the bottom of Table 5-23 ($G=1.63$; $H=2.51$; $r=0.73$) suggest that the six Task A topics can be separated into approximately 2.5 difficulty strata and that the degree of separation between these levels is acceptably high, given the reliability value of 0.73. Moreover, the significant chi-squared statistic implies that these topics are not equal in terms of difficulty. Put differently, when all five criteria are taken into account, the six Task A topics cannot be considered parallel versions of the same task.

In contrast, the topic separation indices in IELTSA portray a different picture (see Table 5-24). Once the TD criterion is removed from the analysis, a drop is observed in the separation, strata and reliability values ($G=0.76$; $\text{Strata}=1.34$, $r=0.36$), indicating that the six topics cannot be separated into statistically distinct difficulty levels and are thus considered parallel; a result also substantiated by the non-significant chi-squared statistic.

5.3.8.2 Task Type B: Topic Measurement Reports

The same analyses as above are now repeated for the four Task Type B topics. The measurement reports from FullA and IELTSA are reproduced in Table 5-25 and Table 5-26 respectively, and sorted in ascending order of topic difficulty measure (in logits). All topic infit and outfit statistics fall within the stringent range of 0.7 to 1.3.

The ranking of topics across the two analyses is not similar. However, Topic B.2 (Describe a River) is identified as the most difficult Task B topic while the remaining three topics have very similar difficulty measures. The observed range in topic difficulty measures is 0.40 and 0.52 logits for FULLA and IELTSA respectively, both of which are considerably larger than the range of difficulty in Task A topics.

Table 5-25 Task Type B: Topic Measurement Report (FullA)

TOPIC	OBSVD AVG	FAIR-M AVG	MEASURE	MODEL S.E	INFIT MNSQ	OUTFIT MNSQ
B.1	5.4	5.64	-0.12	0.07	0.94	1.07
B.3	5.7	5.62	-0.09	0.07	0.95	0.92
B.4	5.6	5.61	-0.07	0.04	0.96	1.29
B.2	5.2	5.4	0.28	0.07	1.17	1.17
Mean (N=4)	5.5	5.57	0	0.06	1	1.11
S.D	0.2	0.11	0.19	0.01	0.11	0.16
Model, Sample: RMSE .08 Adj (True) S.D. .23 Separation 2.93 Strata 4.24 Reliability .90 Model, Fixed (all same) chi-square: 26.8 d.f.: 3 significance (probability): .00						

Table 5-26 Task Type B: Topic Measurement Report (IELTSA)

TOPIC	OBSVD AVG	FAIR-M AVG	MEASURE	MODEL S.E	INFIT MNSQ	OUTFIT MNSQ
B.4	6.1	6.16	-0.15	0.05	0.98	1.05
B.1	5.8	6.15	-0.14	0.08	0.97	1
B.3	6.2	6.12	-0.08	0.09	0.89	0.87
B.2	5.7	5.89	0.37	0.09	1.11	1.11
Mean (N=4)	6	6.08	0	0.08	0.99	1.01
S.D	0.2	0.13	0.25	0.02	0.09	0.1
Model, Sample: RMSE .07 Adj (True) S.D. .18 Separation 2.73 Strata 3.97 Reliability .88 Model, Fixed (all same) chi-square: 21.8 d.f.: 3 significance (probability): .00						

Unlike Task A topics, the results of the two analyses for Task B topics are much more similar in terms of high topic separation indices and associated reliability values ($G_{FullA}=2.93$, $H_{FullA}=4.24$, $r_{FullA}=0.90$ and $G_{IELTSA}=2.73$, $H_{IELTSA}=3.97$, $r_{IELTSA}=0.88$). These results can be taken to suggest that the Task B topics are not equivalent in difficulty (with or without the TD criterion) and that they can be separated into approximately 4 statistically distinct difficulty strata. The significant chi-squared results for both analyses confirm that the task type B topics do not share the same difficulty measures.

An examination of the relative topic difficulty measures illustrate that it is only Topic B.2 which has a markedly different difficulty measure compared to the other three Task B topics. It was therefore hypothesised that this topic might be qualitatively different from

the remaining three topics. The four topics require examinees to describe ‘a friend’ (B.1), ‘a family member’ (B.3), ‘an important choice you had to make’ (B.4) and ‘a river, lake or a sea that you like’ (B.2). The common theme amongst the three first topics is that they appear to be more personal than the latter ‘river’ topic. A possible explanation for the findings is that the test takers were better able to relate to the first three topics compared to the latter.

A final consideration before moving to Task Type C is whether the statistically distinct difficulty strata for Task B topics can have a practical effect on performance scores in the IELTS context. One way of answering this question is to examine the Fair-M Average results for the different topics in IELTSA (see Table 5-26). As previously discussed, the Fair-M Averages are adjustments of the observed averages for any differences in the speaking proficiency of the sample of examinees who responded to those topics and the relative severity of the raters who marked the topics. Topic B.4 has the highest Fair-M average mark of 6.16 while Topic B.2 has the lowest Fair-M average mark of 5.89. The difference between these scores is 0.27, or less than half a band, indicating that the statistical difference in the topic difficulty measures does not translate into a practical influence on the performance scores at the task level. However, these findings are looking at ‘averages’ across criteria. A more in-depth analysis of topic effects on scores is presented at a later section.

5.3.8.3 Task Type C: Topic Measurement Reports

The topic measurement reports for the eight Task Type C topics are reproduced in Table 5-27 and Table 5-28 for FULLA and IELTSA respectively. All topic infit and outfit statistics fall within the stringent range of 0.7 to 1.3.

In FullA the difference between the easiest topic ($T_{C,5}=-0.38$) and most difficult topic ($T_{C,4}=+0.24$) spreads a 0.62 logit range, reduced to 0.54 for IELTSA. The rankings for the easiest and most difficult topics remain the same across the analyses. However, there are differences in the ranking of the topics in between.

Table 5-27 Task Type C: Topic Measurement Report (FullA)

TOPIC	OBSVD AVG	FAIR-M AVG	MEASURE	MODEL S.E	INFIT MNSQ	OUTFIT MNSQ
C.5	5.5	5.55	-0.38	0.04	1.02	1.11
C.7	5.4	5.51	-0.29	0.08	0.91	0.93
C.8	5.6	5.39	-0.08	0.08	1	1.06
C.3	5.2	5.3	0.09	0.07	0.92	0.94
C.1	5.1	5.29	0.1	0.07	0.99	0.98
C.2	5.1	5.27	0.13	0.07	0.94	0.93
C.6	5.4	5.24	0.18	0.07	1.22	1.49
C.4	5.1	5.21	0.24	0.07	0.82	0.82
Mean (N=8)	5.3	5.35	0.00	0.07	0.98	1.03
S.D	0.2	0.12	0.23	0.01	0.12	0.2
Model, Sample: RMSE .07 Adj (True) S.D. .22 Separation 3.05 Strata 4.39 Reliability .90 Model, Fixed (all same) chi-square: 106.2 d.f.: 7 significance (probability): .00						

Table 5-28 Task Type C: Topic Measurement Report (IELTSA)

TOPIC	OBSVD AVG	FAIR-M AVG	MEASURE	MODEL S.E	INFIT MNSQ	OUTFIT MNSQ
C.5	6.1	6.11	-0.34	0.05	1.05	1.07
C.7	6	6.07	-0.26	0.09	0.91	0.89
C.8	6.2	5.98	-0.08	0.09	1.06	1.15
C.6	6	5.95	-0.01	0.09	1.09	1.11
C.3	5.7	5.87	0.15	0.09	0.91	0.92
C.2	5.7	5.86	0.16	0.08	1	0.97
C.1	5.6	5.85	0.18	0.08	0.94	0.94
C.4	5.7	5.84	0.20	0.09	0.81	0.8
Mean (N=8)	5.9	5.94	0.00	0.08	0.97	0.98
S.D	0.2	0.1	0.21	0.01	0.1	0.12
Model, Sample: RMSE .08 Adj (True) S.D. .19 Separation 2.29 Strata 3.38 Reliability .84 Model, Fixed (all same) chi-square: 64.2 d.f.: 7 significance (probability): .00						

Similar to Task Type B topics, an examination of the separation and strata indices and the separation reliability statistics indicate that the eight Task Type C topics can be reliably divided into 3 to 4 statistically distinct difficulty levels ($G_{FullA}=3.05$, $H_{FullA}=4.39$, $r_{FullA}=0.90$; $G_{IELTSA}=2.29$, $H_{IELTSA}=3.38$, $r_{IELTSA}=0.84$). The significant chi-squared results substantiate the finding that regardless of the analysis i.e. with or without the TD criterion, the Task Type C topics cannot be considered parallel versions of the same task. Nevertheless, a consideration of the Fair-M average marks for the eight topics in IELTSA shows that the difference between topics at the two difficulty ends are 0.2 IELTS band scores. As explained earlier, the smallest unit which can have a meaningful difference in IELTS is half a band (0.5). It is therefore unlikely for the differences in difficulty of Task C topics to have a meaningful and practical effect on performance scores at the task level as $0.2 < 0.5$. Nevertheless, it is still possible that differences in topic measures might have an influence at particular score categories.

5.3.8.4 Topic Influence on Scores at the Task Level (FullA)

As explained earlier, a more comprehensive and in-depth approach to considering the impact of topic difficulty on performance scores entails an evaluation of these differences in relation to the scale categories across the different criteria.

Table 5-29 summarises the information extracted from category statistics for each criterion and for the different task types from FullA, where the differences between average ability measures observed at adjacent band scores across criteria were separately calculated. Also included in the table (at the bottom of each section pertaining to a given task type) is the maximum difference between the easiest and most difficult topic

measures within each task type, extracted and calculated from the topic measurement reports for the different task types.

Table 5-29 Increase in Average Ability Measures in Adjacent Categories for All Criteria

TASK TYPE A	BANDS	FC	LR	GA	P	TD (LEVELS)	
	Bands 3-4				1.21		
	Bands 4-5	1.2	1.25	1.34	0.89		
	Bands 5-6	0.89	1.05	0.96	0.75	Levels 1-2	1.08
	Bands 6-7	0.97	0.99	1.07	1.63	Levels 2-3	1.16
	Bands 7-8	1.64	1.32	1.38	0.84	Levels 3-4	0.32
	Bands 8-9	1.22	1.15	1.13	*	Levels 4-5	1.19
	Average	1.18	1.15	1.18	1.06	0.94	
	Task Type A Topics (Maximum Difference in Difficulty)=0.30 logits						
Task Type B	Bands	FC	LR	GA	P	TD (LEVELS)	
	Bands 3-4				0.99		
	Bands 4-5	1.04	1.22	1.36	1.14		
	Bands 5-6	0.84	0.91	0.90	0.49	Levels 1-2	0.97
	Bands 6-7	1.05	0.86	0.92	1.75	Levels 2-3	1.12
	Bands 7-8	1.81	1.81	1.58	1.42	Levels 3-4	0.43
	Bands 8-9	1.53	1.4	1.57		Levels 4-5	0.87
	Average	1.25	1.24	1.27	1.16	0.85	
	Task Type B Topics (Maximum Difference in Difficulty)=0.40 logits						
Task Type C	Bands	FC	LR	GA	P	TD (LEVELS)	
	Bands 3-4				1.02		
	Bands 4-5	0.93	0.97	1.05	0.91		
	Bands 5-6	1.08	1.08	1.02	0.96	Levels 1-2	1.08
	Bands 6-7	0.97	0.97	1.05	2.06	Levels 2-3	1.04
	Bands 7-8	1.81	1.84	1.77	1.3	Levels 3-4	0.58
	Bands 8-9	1.26	0.89	0.96	0.49	Levels 4-5	1.35
	Average	1.21	1.15	1.17	1.12	1.01	
	Task Type C Topics (Maximum Difference in Difficulty)=0.62 logits						

If we look at the data for Task Type A, we see that the maximum difference between the easiest and most difficult topic is 0.30 logits. In order to see whether this value can have a practical (meaningful) impact on performance, the average ability levels required to move across different score categories and for the different criteria are examined. Generally speaking, an increase of approximately one logit is required to move across different bands/levels for the different criteria. Given that $0.30 < 1.00$, it is unlikely for the differences in topic difficulty measures to have a meaningful influence on spoken performance at the task level for Task Type A. The only instance where the 0.30 value

might have an impact is for the TD criterion, where it is shown that an increase of 0.32 logits is required for moving from level 3 to level 4. Given that 0.30 is quite close to 0.32, it can be argued that at this level, an easier topic might facilitate moving to a higher adjacent band or vice versa for the TD criterion.

The same analysis was applied to the remaining task types and similar results emerged: the maximum difference in topic difficulty measures at task types B and C does not exceed the average ability required to move along adjacent score categories for the different criteria. The only instances where topic is likely to have an impact is for the TD criterion where the average ability required to move from level 3 to level 4 is close in value to the corresponding value for the maximum difference in topic measures for each task type. Otherwise, an easier or a more difficult topic is unlikely to have a practical (meaningful) influence – in terms of achieving a higher or lower band score – in relation to the different task types.

5.3.8.5 Topic Influence on Scores at the Task Level (IELTSA)

The above analyses are repeated for IELTSA. Table 5-30 summarises the data which was extracted and calculated from the category statistics for each criterion. The maximum difference between the easiest and most difficult topic measures within each task types are also presented in the same table (this information was extracted from the topic measurement reports for each task type).

Table 5-30 Increase in Average Ability Measures in Adjacent Categories for the IELTS Criteria

TASK TYPE A	BANDS	FC	LR	GA	P
	Bands 3-4				1.38
	Bands 4-5	1.31	1.42	1.44	1.14
	Bands 5-6	1.08	1.23	1.22	1.04
	Bands 6-7	1.25	1.31	1.42	2.26
	Bands 7-8	2.24	1.74	1.88	1.1
	Bands 8-9	1.63	1.63	1.49	1.19
	Average	1.50	1.47	1.49	1.35
	Task Type A Topics (Maximum Difference in Difficulty)= 0.24 logits/0.12 IELTS Bands				
Task Type B	Bands	FC	LR	GA	P
	Bands 3-4				*
	Bands 4-5	1.27	1.31	1.57	1.45
	Bands 5-6	1.07	1.19	1.25	0.83
	Bands 6-7	1.39	1.18	1.2	2.28
	Bands 7-8	2.35	2.27	2.04	1.85
	Bands 8-9	2.04	2.11	2.08	
	Average	1.62	1.61	1.63	1.33
	Task Type B Topics (Maximum Difference in Difficulty)= 0.52 logits/0.27 IELTS Bands				
Task Type C	Bands	FC	LR	GA	P
	Bands 3-4				1.17
	Bands 4-5	1	1.09	1.15	1.16
	Bands 5-6	1.27	1.23	1.22	1.26
	Bands 6-7	1.2	1.25	1.38	2.66
	Bands 7-8	2.37	2.34	2.24	1.71
	Bands 8-9	1.81	1.46	1.45	0.88
	Average	1.53	1.47	1.49	1.47
	Task Type C Topics (Maximum Difference in Difficulty)= 0.54 logits/0.27 IELTS Bands				

The maximum difference between the easiest and most difficult topics at each task type are 0.24, 0.52 and 0.54 for Task Types A, B and C respectively. In terms of the IELTS raw-score metric, these values correspond to 0.12, 0.27 and 0.27 IELTS bands all of which fall below the minimum meaningful unit of 0.5 band in the IELTS scale. Not only have the maximum differences in topic difficulty levels decreased in comparison the previous analysis (FullA), the average ability level required to move along adjacent score categories across the different criteria has also generally increased to approximately 1.5 logits, which

render the likelihood of a topic influence on scores minimal. A scan of the data in the table confirms that this is the case, as the speaking ability required to move along adjacent band scores for all the IELTS criteria consistently exceeds the maximum difference between the easiest and most difficult topics for each task type.

The findings from the above analyses can now be used in answering the research question posed earlier:

When task type is held constant, to what extent are the different topics used in parallel versions of a task similar in terms of difficulty? Are (any) differences in topic difficulty measures large enough to have practical significance in terms of test performance?

In answering the first part of the research question, the topic measurement results and topic separation indices suggest that when all criteria are included in the analyses (FullA), topics within all three tasks types (A, B and C) can be divided into a minimum of two difficulty strata. In other words there are at least two topics within each task type which cannot be considered parallel, as they exhibit difficulty measures which are significantly different. However, in the IELTSA, only the topics in task types B and C were significantly different whereas the six Task Type A topics exhibited very similar difficulty measures and could therefore be considered parallel.

In answering the second part of the research question, the results of the analyses (FullA and IELTSA) suggest that differences in topic difficulty measures (within task types) are not large to have practical significance in terms of achieving a higher or lower band score across different criteria. Broadly speaking, even for task types where at least two of the topics were shown to belong to statistically distinct difficulty strata (e.g. in Task Types B and C), the differences were not large enough to be translated into meaningful differences

in performance scores. The only exception where an easier or more difficult topic can potentially result in the test takers achieving a higher or lower score is for the TD criterion (in IELTSA) but the effect is only likely to be observed at specific adjacent categories (levels 3 and 4).

The influence of topic on performance within each task type remains the theme of the next section. However, the focus is shifted from an examination of topic effects in relation to scores to an examination of the functions elicited by different topics across the three task types. The analyses of the functions do not directly follow the MFRM analyses. Nevertheless, they are reported in succession, as they can provide complementary perspectives on the effects of topic on performance.

5.3.9 Functions Analyses

Another way of looking at the influence of topic on performance is to consider the functions elicited from topics within each task type. This focus on functions reflects the purpose of the IELTS Speaking Tests in eliciting different types of functions from test takers. This section aims to answer the following research question:

- v. When task type is held constant, to what extent are the observed functions elicited by different topics similar?*

To remind the reader, raters in the study were asked to complete an ‘observation checklist’; an instrument which allows for a comparison of different speaking tasks in terms of the range of functions they can elicit. The data from the observation checklist were sorted by task type so that topics within each task type could be compared in terms of the range of elicited functions. It should be noted that raters identified *types* of

functions observed in each performance and not the *frequency* with which each function was observed⁴⁶.

Due to space limitations, the full results of the functions analyses are not presented here. However, illustrative examples are provided in Table 5-31.

The first column in the table lists the different types of functions in the observation checklist. In the remaining columns, two topics from each task type were selected where one topic is amongst the easiest and the other topic is amongst the most difficult of the topics within each task type on the basis of the MFRM results. The (+) sign next to the topic reference number represents the easier topic, compared to the more difficult topic represented with a (-) sign. The description for each topic is provided at the bottom of the table.

It should be noted that the data in the table are expressed in percentages i.e. the percentage of times a specific function was observed (at least once) in a specific task. For example, of the total number of times Topic A.6 (Dancing) was responded to, the ‘personal information function’ was observed at least once 87.2% of the time. In contrast, the same function was only observed 11.9% of the time for Topic A.3 (Festivals). In fact, amongst all Task Type A topics, Topic A.3 had the smallest percentage of observations for personal information. Given that Task A topics are designed to be ‘familiar’ topics, the lack of observations for the personal information function can be indicative of the impersonal and/or non-familiar nature of the topic. In light of the MFRM results, this finding can also explain why Topic A.3 (designed to be familiar) exhibited a difficulty

⁴⁶ An exact agreement of 79.2% was reached for the type of functions observed in the common batch performances.

measure closer to Task C topics (designed to be more abstract and arguably more unfamiliar).

Table 5-31 Percentage of Observed Functions by Task Type (Illustrative Examples)

TASK TYPE	A		B		C	
FUNCTION	A.6 (+)	A.3 (-)	B.1 (+)	B.2 (-)	C.5 (+)	C.6 (-)
Personal Information	87.2	11.9	93.5	33.3	43.6	12.8
Non-Personal Information	76.9	90.5	17.4	61.9	41	89.7
Expressing Opinions	79.5	78.6	32.6	57.1	97.4	89.7
Justifying Opinions	66.7	59.5	19.6	42.9	82.1	74.4
Explaining	41	42.9	43.5	26.2	35.9	25.6
Suggesting	0	9.5	0	2.4	12.8	5.1
Expressing Preferences	20.5	11.9	4.3	40.5	0	2.6
Comparing	12.8	11.9	17.4	16.7	46.2	17.9
Contrasting	15.4	21.4	15.2	9.5	28.2	10.3
Paraphrasing	17.9	11.9	13	7.1	12.8	10.3
Narrating	12.8	11.9	39.1	23.8	2.6	0
Describing	23.1	69	63	81	5.1	2.6
Elaborating	46.2	45.2	47.8	38.1	53.8	35.9
Summarising	15.4	11.9	8.7	4.8	15.4	10.3
Speculating	25.6	7.1	0	2.4	15.4	23.1
Staging	2.6	4.8	4.3	4.8	17.9	10.3
Analyzing	2.6	4.8	8.7	0	10.3	2.6
Negotiation of Meaning	10.3	9.5	2.2	7.1	15.4	12.8
Conversation Repair	0	2.4	0	0	0	0
Self Repair	33.3	19	32.6	19	20.5	28.2
Agreeing	5.1	16.7	2.2	0	5.1	5.1
Disagreeing	0	2.4	0	0	30.8	2.6
Commenting on Topic Difficulty	12.8	33.3	0	7.1	12.8	38.5
Task Topic Clarification Request	7.7	21.4	4.3	2.4	15.4	20.5

A.6= Dancing; A.3= Festivals; B.1 =Describe a Friend; B.2= Describe a River; C.5=Family Similarities; C.6= Genetic Research

Other large differences were observed for the ‘describing’ function with a higher percentage of observation (69%) for Topic A.3 compared to A.6 (23.1%). Another striking contrast is observed for the two topic-specific functions of ‘commenting on topic difficulty’ and ‘task topic clarification request’ both of which were markedly higher for A.3 compared to A.6. Lastly, the ‘speculating’ function was observed 25.6% of the time

for A.6 whereas it was only observed 7.1% of the time for A.3. In fact, the speculating function was generally observed about 5-11% of the time for the Task A topics, with the exception of A.6. In terms of the other functions, the percentages appear to be more similar. This is particularly the case for functions such as justifying opinions, expressing opinions, explaining, comparing and contrasting and elaborating.

If we now consider the two task B topics i.e. Topic B.1 (Describe a Friend) and B.2 (Describe a River) and focus on the proportion of personal to non-personal functions, we see a sharp drop in the percentage of personal information function from 93.5% for Topic B.1 to 33.3% for Topic B.2. The reverse pattern is detected for the non-personal information function where there is a dramatic increase from 17.4% in Topic B.1 to 61.9% in Topic B.2. The same pattern is repeated for the two Task Type C topics. It therefore appears that easier topics are associated with higher percentage of personal-information functions whereas more difficult topics are associated with lower personal-information functions. An indirect inference can thus be made that topics which have been shown to be psychometrically easier for test takers (on the basis of MFRM results), seem to require test takers to draw more extensively on personal information which is arguably more readily available to them.

Other large differences between the percentage of observed functions in Task B topics are for functions such as 'expressing' and 'justifying opinions' and most markedly in 'expressing a preference' which appear more in B.2 compared to B.1. On the other hand, 'narrating' is more often observed in Topic B.1 (39.1%) vs. Topic B.2 (23.8%). Lastly, the 'commenting on topic difficulty' function was not observed in Topic B.1 although it was observed for Topic B.2 about 7% of the time, which can be used as confirmatory evidence

for the MFRM analyses in which B.2 was found to be more difficult for participants compared to B.1.

Finally, if we examine the percentages of observed functions in Topics C.5 (Family Similarities) and C.6 (Genetic Research) from Task Type C, we can see that (apart from the personal and non-personal functions discussed earlier) there are some noticeable differences for the ‘compare and contrast’ functions which are more frequently observed in Topic C.5 compared to C.6. On the other hand, the ‘disagreeing’ function is most noticeably observed for Topic C.5 (30.8%) compared to C.6 (2.6%). In fact, an examination of the percentage of the occurrence of this function across topics and across task types reveals that the disagree function is generally rarely observed and when it is observed, its percentage is limited to between 2-8%. Topic C.5 is the only exception with a substantively higher percentage of observation. In terms of the ‘comment on topic difficulty’ function, once again a similar pattern to the previous task types is observed, where a higher percentage of this function is exhibited for the more difficult topic.

A more in-depth analysis of the distribution of functions across all task types and topics is not possible due to space limitations. However, the examples provided above serve to illustrate that topics within a task type may not necessarily elicit similar functions. Put differently, topics may not have an influence on performance scores but they can nevertheless have an influence on the range of functions elicited. In answering the research question, the analyses have shown that when task type is held constant, there are both similarities and differences in terms of the functions the topics elicit.

Before moving on, attention is drawn to Table 5-32, where the distribution of the *average* percentage of observations for each function is provided across all speaking tasks. Functions which have similar average values are identified with the same colour.

Table 5-32 Average Percentage of Observed Functions Across All Tasks

FUNCTION	Avg (%)
Expressing Opinions	66.4
Providing Personal Information	57.2
Providing Non-Personal Information	52.0
Justifying Opinions	50.0
Elaborating	47.1
Explaining	38.6
Describing	34.2
Self-Repair	30.5
Comparing	26.9
Expressing Preferences	24.1
Contrasting	23.4
Summarising	15.3
Narrating	12.6
Task Topic Clarification Request	12.4
Paraphrasing	12.3
Comments on Topic Difficulty	10.4
Speculating	10.2
Negotiation of Meaning	9.1
Staging	8.2
Suggesting	5.0
Agreeing	4.8
Analyzing	4.7
Disagreeing	2.7
Conversation Repair	0.2

The information in the table reflects the information-oriented nature of the IELTS Speaking Tests as evidenced in the dominance of the informational functions (e.g. expressing opinions, providing personal and non-personal information, justifying opinions, comparing and contrasting) which appear at the top of the table and can be distinguished from interactional functions (e.g. negotiation of meaning, agreeing, disagreeing) which appear at the bottom of the table and where their associated average

percentages suggest rare observations. Combined with the striking 0.2% of observation for the ‘conversation repair’ function, the extent to which the speaking tests are successful in eliciting interactional functions can be brought into question. Furthermore, these findings highlight the central role of *information* in these tests as the base around which performance is built.

Having discussed the influence of topic on performance scores on the one hand, and on functions elicited on the other, we now return our attention to the final output from the first MFRM analysis.

5.3.10 Global Model Fit

The analyses of the specified facets of assessment in the current study, which included an examination of the main effects of examinee ability, rater severity, criteria difficulty, rating scale functionality and speaking task/topic difficulty, has been completed. While the main purpose of the analyses was to answer the topic-related research questions, the analyses of other facets were given equal weight, as on the one hand, they serve as a quality control check for the various aspects of the speaking tests used in the study, and on the other, they provide a more comprehensive picture in which the influence of topics can be understood. However, before moving on to the next analysis, the global fit of the data to the Rasch model has to be first addressed.

As mentioned earlier, in examining the global fit of data to the Rasch model, it is important to note that misfit to the Rasch model is deviation from perfection and that it is ‘practical utility’ of the model which should be borne in mind (Eckes, 2009:27). A useful way of evaluating the overall fit of the data to the model is to examine another output from FACETS: the table of unexpected responses. According to Linacre (2011a), model fit can

be considered acceptable when the percentage of standardised residuals which fall outside of the range of -2.00 to 2.00 are 5% or less and the percentage of standardised residuals which fall outside of the criterion range of -3.00 to +3.00 are 1% or less.

Considering the present sample, the results are as follows:

- FullA : Total number of valid responses=7010; Standardised residuals $\geq |2|$ = 100 and standardised residuals $\geq |3|$ = 41 i.e. 1.43% $\geq |2|$ and 0.58% $\geq |3|$.
- IELTSA: Total number of valid responses=5608; Standardised residuals $\geq |2|$ = 100 and standardised residuals $\geq |3|$ = 23 i.e. 1.78% $\geq |2|$ and 0.41% $\geq |3|$.
- Overall, these findings suggest a satisfactory fit of the data to the Rasch model.

The table of unexpected responses can also be scanned for trends which can provide valuable insight into aspects of the assessment context. An examination of the table of unexpected responses for FullA revealed a striking pattern where an overwhelming majority of the highest standardised residuals were associated with the TD criterion (see Appendix S). In particular, unexpected ratings were found for persons of high speaking ability measures for whom the Rasch model predicted high TD scores whereas the observations showed very low scores, resulting in large standardised residuals. These findings explain why the TD criterion exhibited larger infit statistics compared to the other criteria. The unexpected performances pertaining to the TD criterion will be examined more closely in the next chapter.

5.4 Five-Facet MFRM Analysis

Earlier in the chapter and in relation to the conceptual-psychometric model (Eckes, 2009) adopted in the study, it was explained that test taker characteristics (such as gender) are normally considered as distal facets and are examined in interaction/bias analyses. However, background knowledge of topic (BK), unlike gender, is not a stable characteristic of the test taker and might fluctuate from one topic to the next. Given that one of the study's main research questions aims to evaluate the main effect of BK on performance, BK was conceptualised as a proximal facet of assessment and a new MFRM analysis was run where the following proximal facets and elements were specified:

- Examinee facet (81 participant elements)
- Rater facet (4 rater elements)
- Rating Scale (5 criteria elements)
- Speaking Tasks (18 topic elements)
- Background Knowledge of Topic (3 condition elements)

The BK estimates for each person x topic combination (derived from the analysis of the BK questionnaires, discussed in the preliminary analyses chapter) were divided into three groups of Low, Medium and High constituting the three elements of the BK facet. The reason why these BK estimates are grouped is because FACETS only accepts integer numbers. The FACETS manual therefore suggests 'chunking this [continuous data] up into qualitatively advancing pieces and number the chunks' (Linacre, 2011a: 247).

Similar to the previous section, separate MFRM analyses are carried out with and without the TD criterion. However, the focus of next sections is predominantly on the BK and topic facets, as the other facets were thoroughly discussed in the previous analysis.

Nevertheless, the topic, rater and criteria measurement reports from the analyses with and without BK will be compared against each other in order to examine whether adding BK as a facet has exerted any influence on the estimates of the other facets.

5.4.1 Facet Map

The vertical maps in Figure 5-8 and Figure 5-9 are a visual representation of the MFRM analysis for all five criteria (FullA) and the analysis without TD (IELTSA) respectively. The BK facet now appears in the fifth column of the diagrams.

Ability (High)		Severe	Difficult			FC	LR	GA	PR	TD
Measr	+examinee	-Rater	-Topic	-BKGroup	-Cri	S.1	S.2	S.3	S.4	S.5
4	+	+	+	+	+	(9)	(9)	(9)	(9)	(5)
	*								8	
	*									
3	+	+	+	+	+	8	8	8		
	*								---	---
									7	
2	+	+	+	+	+					
	**									
	**					---	---	---	---	4
	*									
1	+	+	+	+	+					
	*									
	*									
	***							7		---
	***	R2	C.8	LowBK	P	7	7		6	
:	:	R4	C.1		FC					
:	:		C.2		GA					
:	:		B.2		LR					
:	:		C.4							
*	0	*	C.6	* MediumBK	*	*	*	*	*	3
:	:		C.7							
:	:		A.3							
:	:		A.4							
:	:		A.5							
:	:		A.1							
:	:		A.2							
:	:		B.1							
:	:		C.3							
:	:		B.3							
:	:		C.5	HighBK				---		
:	:		B.4							
:	:		A.6							
:	:									
	*****					---	---		---	---
	*****	R3								
-1	+	+	+	+	+					
	*****					6	6	6		
	*****				TD				5	

	**									2
-2	+	+	+	+	+		---	---		
	**									

	*								---	
	*									---
-3	+	+	+	+	+	5	5	5		
	*									
	*									
	*								4	
-4	+	+	+	+	+	(4)	(4)	(4)	(3)	(1)
Measr	* = 1	-Rater	-Topic	-BKGroup	-Cri	S.1	S.2	S.3	S.4	S.5
Ability (Low)		Lenient		Easy		FC	LR	GA	PR	TD
Mean	-0.63	0.00	0.00	0.00	0.00					
S.D	1.37	0.47	0.16	0.34	0.67					

Note: Each start (*) in the second column represents 1 examinee.
FC=Fluency and Coherence, LR=Lexical Resource, GA=Grammatical Accuracy, PR=Pronunciation, TD=Topic Development.

Figure 5-8 Facet Map (5-Facet, FullA)

Ability (High)		Severe	Difficult			FC	LR	GA	PR
Measr	+examinee	-Rater	-Topic	-BKGroup	-Cri	S.1	S.2	S.3	S.4
5	+	+	+	+	+	(9)	(9)	(9)	(9)
4	**	+	+	+	+	8	8	8	8
	*								---
3	+	+	+	+	+	+	+	+	+
	*								7
2	**	+	+	+	+	+	+	---	+
	*					---	---		---
1	+	+	+	+	+	+	+	+	+
	*	R2				7	7	7	
	**	R4	C.8	LowBK	P				6
:	:	R1	C.1		:	:	:	:	:
:	:		C.2		:	:	:	:	:
:	:		B.2		:	:	:	:	:
:	:		C.3		:	:	:	:	:
:	:		B.3		:	:	:	:	:
*	0	*	C.5	MediumBK	* GA	*	*	*	*
:	:	:	C.6		:	:	:	:	:
:	:	:	C.7		:	:	:	:	:
:	:	:	A.4		:	:	:	:	:
:	:	:	A.1		:	:	:	:	:
:	:	:	A.2		:	:	:	:	:
:	:	:	B.1		:	:	:	:	:
:	:	:	C.4		:	:	:	:	:
:	*****	:	B.4	HighBK	FC	:	:	:	:
:	:	:	A.3		LR	:	:	:	:
:	:	:	A.5		:	:	:	:	:
:	*****	:	A.6		:	:	:	:	:
-1	+	R3	+	+	+	---	---	---	---
	*****						6	6	
	*****					6			5
-2	+	+	+	+	+	---	---	---	---

	*								
-3	+	+	+	+	+	5	5	5	---

	**								
	**								
-4	+	+	+	+	+	---	---	---	4
	**								
-5	+	+	+	+	+	(4)	(4)	(4)	(3)
Measr	* = 1	-Rater	-Topic	-BKGroup	-Cri	S.1	S.2	S.3	S.4
Ability (Low)		Lenient		Easy		FC	LR	GA	PR
Mean	-1.18	0.00	0.00	0.00	0.00				
S.D	1.73	0.73	0.18	0.32	0.25				

Note: Each start (*) in the second column represents 1 examinee.
FC=Fluency and Coherence, LR=Lexical Resource, GA=Grammatical Accuracy, PR=Pronunciation

Figure 5-9 Facet Map (5-Facet, IELTSA)

5.4.2 The Background Knowledge Facet

The literature suggests that higher levels of BK may have a facilitative effect on performance. On this basis, we would assume the LowBK condition to be the most challenging condition for test takers and therefore located towards the top (difficulty) end of the facet map while the HighBK condition would appear towards the bottom easy end of the logit scale. On the other hand, if BK does not have an influence on performance then the different BK conditions would not appear in any particular order and their measures would be very close in difficulty. The BK separation statistics would also be small with a reliability value of close to 0.

A glimpse at the facet maps indicates that not only are the different BK conditions ordered as predicted, that is, from high to low in ascending order of difficulty, but that there is a notable distance between the BK element measures. This preliminary observation of the data is indicative of the facilitative effect of higher level BK on performance and thus warrants further analysis. The BK measurement report for FULLA and IELTSA are presented in Table 5-33 and Table 5-34 respectively.

Table 5-33 The Background Knowledge Measurement Report (FullA)

BK LEVEL	OBSVD AVG	FAIR- M AVG	MEAS URE	MODEL S.E	INFIT MNSQ	OUTFIT MNSQ
High BK	5.7	5.63	-0.32	0.03	1.09	1.23
Medium BK	5.3	5.48	-0.05	0.03	0.9	0.93
Low BK	5.3	5.24	0.36	0.03	1	1.09
Mean (N=3)	5.4	5.45	0.00	0.03	1	1.08
S.D	0.2	0.2	0.34	0.00	0.09	0.15
Model, Sample:RMSE .03 Adj (True) S.D .34 Separation 12.13 Strata 16.51 Reliability .99 Model, Fixed (all same) chi-square: 285.2 d.f.: 2 significance (probability): .00						

Note: BK= Background Knowledge

Table 5-34 The Background Knowledge Measurement Report (IELTSA)

BK LEVEL	OBSVD AVG	FAIR- M AVG	MEAS URE	MODEL S.E	INFIT MNSQ	OUTFIT MNSQ
High BK	6.2	6.17	-0.29	0.04	1.1	1.12
Medium BK	5.9	6.05	-0.05	0.03	0.92	0.92
Low BK	5.9	5.85	0.34	0.03	0.95	0.97
Mean (N=3)	6	6.03	0	0.03	0.99	1
S.D	0.2	0.16	0.32	0	0.09	0.1
Model, Sample: RMSE .03 Adj (True) S.D. .32 Separation 9.33 Strata 12.78 Reliability .99 Model, Fixed (all same) chi-square: 169.6 d.f.: 2 significance (probability): .00						

Note: BK= Background Knowledge

The BK estimates in the tables are arranged in ascending difficulty measure (in logits) and illustrate that the High BK condition is associated with the lowest measures (HighBK_{FullA}=-0.32; HighBK_{IELTSA}=-0.29) whereas the Low BK condition has the highest measures (LowBK_{FullA}=+0.36; LowBK_{IELTSA}=+0.34) spanning a similar range of 0.68 and 0.63 across the two analyses. The MediumBK estimate remains the same in FullA and IELTSA. The infit mean square statistics for all elements fall within the acceptable control limits of 0.5 to 1.5. In fact, the values range between 0.90-1.1 and are therefore very close to their expected value of 1.0.

The difficulty span suggests that the BK conditions are different from another in terms of the challenge they pose for the test takers. The group statistics at the bottom of the table are examined next in order to evaluate the extent to which the BK conditions are different.

To remind the reader, the interpretation of the separation indices is dependent on the facet under investigation. In terms of the BK facet, the separation index is an indication of the number of statistically distinct difficulty levels that the BK conditions can be divided into and the extent to which the measures are different. Just as we do not want the relative severity of raters to introduce measurement error to an assessment context, we also do not

want BK condition (an arguably construct irrelevant factor) to exert a significant influence on performance. It is therefore expected for the BK separation indices to be very low and the separation reliability value to be close to 0. However, the results from the tables show the exact opposite:

The separation indices for FullA ($G_{\text{FullA}} = 12.13$ $H_{\text{FullA}} = 16.51$ $r_{\text{FullA}} = 0.99$) suggest that BK conditions can be reliably separated into approximately 16 statistically distinct difficulty strata with a high degree of separation between levels as evidenced in the high reliability value of 0.99. In IELTSA the separation index and difficulty strata are comparatively lower ($G_{\text{IELTSA}} = 9.33$ $H_{\text{IELTSA}} = 12.78$ $r_{\text{IELTSA}} = 0.99$) but nevertheless significantly high. Moreover, the null hypothesis that these measures are the same is rejected on the basis of the significant chi-squared statistics ($\chi^2_{\text{FullA}} = 285.2$; $\chi^2_{\text{IELTSA}} = 169.6$) and the corresponding probability values $p = 0.00 < 0.01$.

The above results confirm that the BK condition can have a significant impact on performance. What remains to be addressed is whether or not this influence has practical significance in terms of performance scores. This can be achieved by considering the Fair-M average column in Table 5-34 which is an adjustment of the observed average for the differences in the speaking proficiency of examinees, relative severity of raters and difficulty of topics. The difference between the Fair-M average results of the High BK and Low BK conditions is 0.3, falling beneath the minimum IELTS band unit (0.5) which can have a practical and meaningful effect on candidate performance scores.

Another approach (previously exercised in looking at topic influence on scores) is to consider the effects of BK in relation to average examinee ability levels necessary to move across score categories in the different criteria. On the basis of the category statistics

measurement report, these average measures were calculated and are presented in Table 5-35 and Table 5-36 for FullA and IELTSA respectively.

Table 5-35 Increase in Average Ability Measures in Adjacent Categories for All Criteria

BANDS	FC	LR	GA	P⁴⁷	TD (LEVELS)	
Bands 3-4				0.74		
Bands 4-5	1.04	1.07	1.14	0.88		
Bands 5-6	0.91	0.97	0.95	0.76	Levels 1-2	1.05
Bands 6-7	0.94	0.92	0.94	1.83	Levels 2-3	1.04
Bands 7-8	1.7	1.54	1.6	1.04	Levels 3-4	0.47
Bands 8-9	1.32	1.23	1.07	0.77	Levels 4-5	0.98
Average	1.18	1.15	1.14	1.06	0.88	

Table 5-36 Increase in Average Ability Measures in Adjacent Categories for the IELTS Criteria

BANDS	FC	LR	GA	P⁴⁸
Bands 3-4				0.89
Bands 4-5	1.13	1.14	1.22	1.13
Bands 5-6	1.11	1.17	1.21	1.05
Bands 6-7	1.18	1.18	1.23	2.4
Bands 7-8	2.25	1.98	2.06	1.38
Bands 8-9	1.78	1.76	1.52	1.2
Average	1.49	1.45	1.45	1.43

The maximum difference between the most difficult (Low BK) and the easiest (High BK) BK conditions was calculated at 0.68 and 0.63 in FullA and IELTSA respectively. A close examination of the tables reveals that the minimum average ability required to move along adjacent band levels consistently exceeds the maximum difference between the lowest and highest BK condition measures, making it unlikely for BK to have a practical effect on performance scores. Once again, the only exception appears to be for the TD criterion

⁴⁷ Due to the problems with reversed thresholds for the Pronunciation criterion, the observed average value was replaced with the expected value for the disordered categories.

⁴⁸ Due to the problems with reversed thresholds for the Pronunciation criterion, the observed average value was replaced with the expected value for the disordered categories.

where the ability required to move between levels 3 and 4 (0.47) is smaller than 0.68 and implies that at this particular score threshold, an examinee with a higher BK might be assigned a higher TD level and vice versa.

Taken together, the above results can answer the following research questions:

- vi. Will the differences in test takers' levels of background knowledge of topics have an impact on performance? Are (any) differences large enough to have practical significance in terms of test performance?*

The findings suggest that different levels of that BK can pose significantly distinct levels of challenge for test takers. However, despite this statistical significance, the differences in BK failed to translate into practical significance in terms of impact on performance scores. The maximum difference between the BK conditions was consistently lower than the minimum speaking ability required to move across adjacent bands for the different criteria. This trend was similar across both FullA and IELTSA. The only exception was observed for the TD rating scale (in FullA) for which BK can potentially exert an influence in achieving a higher or lower score at levels 3 and 4.

The above analyses have shown that, at least from a modeling perspective, BK can be explicitly parameterized as an additional facet in MFRM. A question that one can pose in relation to the finding that different levels of BK can pose significantly distinct levels of challenge for test takers, is whether the Rasch model *should* account for these BK differences in the same way, for example, that relative rater severity is accounted for? It is argued here that such an approach is problematic on two grounds. Firstly, from a practical standpoint, it is difficult, if not impossible, to elicit each test taker's background knowledge of different topics in operational test settings. Secondly, from a conceptual

standpoint, by explicitly parameterising the BK facet, the model would be adjusting the examinee raw scores by penalizing those examinees who happen to have high BK of topics while rewarding those who do not! In other words, despite its psychometric value, the approach does not hold water conceptually.

Having examined the effects of BK on performances scores, the influence of BK as an additional facet is now examined in relation to the other facets' estimates to see whether there are any noticeable changes in their measurement results. To this end, Table 5-37 and Table 5-38 provide a comparison of the Rater, Topic, Criterion and Examinee measurement results without the BK facet (on the left) and with the BK facet (on the right) for FullA and IELTSA respectively.

The rater facet results show the exact same ranking of raters with very small differences (0.01-0.04 logits) in rater severity estimates. These differences are even more negligible for the criteria estimates (0.01-0.02). In other words the element measures for these two facets are stable regardless of the inclusion of BK as a facet.

The measurement results for all examinee elements could not be provided in the table. Instead, the mean, standard deviation (S.D) and ability range for each analysis are presented. In both FullA and IELTSA, the mean and S.D of examinee ability has remained stable across the MFRM runs with and without BK. However, the range of examinee ability has stretched from 6.81 (without BK) to 7.08 (with BK) in FULLA. Likewise, an increase of 0.13 from 8.49 to 8.62 is observed for IELTSA.

Table 5-37 MFRM Measurement Result Comparisons (FullA)

	MFRM WITHOUT BK			MFRM WITH BK		
Rater	Facet Elements	Measure	Model S.E.	Facet Elements	Measure	Model S.E.
	R3	-0.65	0.03	R3	-0.69	0.03
	R1	0.09	0.03	R1	0.11	0.03
	R4	0.22	0.03	R4	0.22	0.03
	R2	0.34	0.03	R2	0.35	0.03
Topic	A.6	-0.31	0.07	A.6	-0.35	0.07
	B.4	-0.29	0.08	B.4	-0.27	0.08
	B.1	-0.27	0.07	C.5	-0.15	0.07
	B.3	-0.23	0.07	B.1	-0.09	0.07
	A.1	-0.21	0.05	A.3	-0.09	0.07
	A.2	-0.15	0.05	A.1	-0.08	0.05
	A.5	-0.13	0.07	A.5	-0.07	0.07
	C.5	-0.06	0.07	A.2	-0.05	0.05
	C.7	-0.01	0.07	B.3	-0.03	0.07
	A.4	0.03	0.07	C.7	0.03	0.07
	A.3	0.06	0.07	A.4	0.04	0.07
	B.2	0.12	0.07	C.6	0.11	0.07
	C.8	0.16	0.08	C.3	0.12	0.07
	C.1	0.18	0.07	C.8	0.14	0.08
	C.3	0.19	0.07	B.2	0.15	0.07
	C.2	0.22	0.07	C.2	0.16	0.07
	C.4	0.32	0.07	C.4	0.17	0.07
	C.6	0.39	0.07	C.1	0.26	0.07
Criteria	TD	-1.15	0.04	TD	-1.17	0.04
	FC	0.14	0.04	P	0.14	0.04
	LR	0.18	0.04	FC	0.18	0.03
	GA	0.36	0.04	LR	0.37	0.03
	P	0.47	0.04	GA	0.48	0.03
Examinee	Examinee (Mean)	Examinee (S.D)	Ability Range	Examinee (Mean)	Examinee (S.D)	Ability Range
	-0.63	1.36	6.81	-0.63	1.37	7.08

Table 5-38 MFRM Measurement Result Comparisons (IELTSA)

	MFRM WITHOUT BK			MFRM WITH BK		
Rater	Facet Elements	Measure	Model S.E.	Facet Elements	Measure	Model S.E.
	R3	-1.06	0.03	R3	-1.09	0.03
	R1	0.30	0.03	R1	0.30	0.03
	R4	0.31	0.03	R4	0.34	0.03
	R2	0.45	0.03	R2	0.46	0.03
Topic	A.6	-0.37	0.09	A.6	-0.4	0.09
	A.5	-0.27	0.09	A.5	-0.21	0.09
	A.1	-0.23	0.06	A.3	-0.14	0.09
	B.1	-0.23	0.08	B.4	-0.13	0.09
	A.2	-0.19	0.06	A.1	-0.11	0.06
	B.4	-0.15	0.09	A.2	-0.09	0.06
	B.3	-0.05	0.09	C.5	-0.08	0.09
	A.4	-0.05	0.09	B.1	-0.07	0.08
	A.3	0	0.09	C.6	-0.04	0.09
	C.5	0	0.09	A.4	-0.03	0.09
	C.7	0.04	0.09	C.7	0.07	0.09
	C.2	0.19	0.08	C.4	0.1	0.09
	C.8	0.2	0.09	C.2	0.13	0.08
	C.3	0.2	0.09	C.3	0.14	0.09
	C.1	0.21	0.08	B.3	0.14	0.09
	C.6	0.23	0.09	C.8	0.17	0.09
	B.2	0.23	0.09	B.2	0.27	0.09
	C.4	0.25	0.09	C.1	0.28	0.08
Criteria	FC	-0.21	0.04	FC	-0.22	0.04
	LR	-0.17	0.04	LR	-0.17	0.04
	GA	0.07	0.04	GA	0.07	0.04
	P	0.32	0.04	P	0.32	0.04
Examinee	Examinee (Mean)	Examinee (S.D)	Ability Range	Examinee (Mean)	Examinee (S.D)	Ability Range
	-1.18	1.72	8.49	-1.18	1.73	8.62

When the differences between examinee measurement estimates with and without BK were calculated and examined for individual examinees, the maximum absolute difference was shown to be 0.35 logits for FullA and 0.32 logits for IELTSA (approximately 0.2 IELTS bands). These differences are therefore not large enough to be translated into meaningful differences on scores ($0.2 < 0.5$ IELTS band).

An examination of the topic facet reveals a very different pattern compared to the other facets. First of all, the ranking of the topics has not remained the same in analyses with and without BK. In FullA, the range of topic difficulty is 0.7 (without BK) but the range is reduced to 0.61 when BK is included. In contrast, in IELTSA, the range of topic difficulty has increased from 0.62 logits (without BK) to 0.68 (with BK).

Furthermore, relative to the other facets, the differences in the measurement results of topics are more pronounced, ranging from 0.01 to 0.28 (FullA) and from 0.02 to 0.27 (IELTSA). For example, when BK is included, striking differences are observed for Topics C.6 (Genetic Research) and A.3 (Festivals) which show a drop of 0.28, 0.15 in their difficulty measures respectively. If we look back to the BK questionnaire results (see Table 4-8) we can see that examinees had reported their lowest BK for these same topics. On the other hand, the easier topics (associated with higher BK) show an increase in their difficulty estimates as observed for Topic A.1 (Family) and Topic B.1 (Describe a Friend). FACETS has therefore adjusted topic difficulty measures for differences in test taker characteristics of background knowledge of topics.

These above findings can be used to answer the following research question:

vii. Does background knowledge of topics have an impact on topic difficulty measures?

When examining the influence of BK on the measurement results of the other facets, it was observed that BK had virtually no impact on the measurement results of raters and criteria. A far more pronounced effect was observed for the topic facet in terms of ranking of elements, measurement estimates and difficulty range. In other words, unlike the other facets, the speaking task/topic measurement results fluctuated as a function of the test taker characteristic of background knowledge of topic and did not remain stable. This is a very important finding, as it brings into question the possibility of objectively measuring prompt/task/topic difficulty. The study has shown that difficulty of topics can be influenced by the background knowledge that test takers bring to the test and therefore is not necessarily inherent to the task. Such an observation was most palpable for topics such as Genetic Research (C.6) and Festivals (A.3). These topics will be re-visited in the next chapter where the qualitative results can help shed light on aspects of these topic which might have rendered them more difficult for test takers.

In interpreting the results of the MFRM analyses, an important caveat, which is reflective of the design of the study, should be borne in mind. It was explained in the methodology chapter that not all participants responded to all topics because of practical considerations on the one hand and the desire to include as many topics as possible on the other. A linked design was therefore adopted both in the collection of data (through two common topics) and in the rating design (through a common batch system), thus satisfying the requirements for a connected data set. The Rasch model is considered to be very robust in dealing with missing data (Eckes, 2009), as the model looks at observed patterns in the data and on the basis of the available information, calculates estimations for the unobserved data as well as the observed data. To illustrate, if an examinee is found to

demonstrate high scores on five moderately difficult speaking tasks, then it is likely that the examinee would also perform well on a sixth speaking task with a similar difficulty level, even though in the data collection stage the examinee did not respond to that particular task. The same is true for rater severity. However, a serious problem arises when BK is considered. The participants in the study did not fill out the BK questionnaire for all topics, but only for the ones they responded to, thus resulting in missing BK data. The Rasch model response to the missing data is to look for patterns in BK groupings and to make estimates about test takers' BK groupings on the basis of available data. For example, if the persons who responded to Topic C.6 (Genetic Research) generally fell into the low BK group, then the model assumes that a person who has not responded to Topic C.5 is likely to also fall in the low BK group. This assumption can be brought into question in light of the highly individual and test-taker dependent nature of BK. Nevertheless, this assumption was necessary for running the analysis for the whole sample and for illustrative purposes. In the next section this limitation is addressed by adopting a different approach, where only the available BK data is used in the analyses and therefore, the need to make any assumptions about the level of background knowledge is circumvented.

5.5 Multiple Regression Analysis: Background Knowledge as a Predictor Variable

The missing BK data was identified as an important limitation of the previous MFRM analysis. Moreover, BK measures could not be directly used in the analyses and had to be grouped into levels, as FACETS only accepts integer numbers. In addressing both these

issues, a decision was made to use a different statistical technique, that is, multiple regression to examine the extent to which different predictor variables can predict spoken performance. This approach not only allows for different Rasch-calibrated measures to be directly used in the analysis but also allows for a comparison of results against other studies in the literature which have used more traditional statistical techniques.

To remind the reader, for every person x topic combination, a separate background knowledge measure was calculated (see results of BK analyses in section 4.2.3). In order to run the analysis, the next step was to estimate a separate speaking ability measure for every person x topic combination as well, so that the influence of background knowledge of a given topic could be directly examined in relation to performance on that specific topic. So far, examinee ability measures were based on the results of examinees responding to all 10 topics. In order to estimate examinee measures on the basis of just one topic, the data first had to be re-arranged so that each person x topic combination would be treated as a distinct person. Next, a 3-facet MFRM analysis (examinee, rater and criteria facets) was run and person x topic speaking measures were estimated. In this approach, topic is no longer considered a separate facet; instead, relative topic difficulties are absorbed in the resulting person x topic measures. To illustrate, a hypothetical example of such a data set for one examinee (Z) is presented in Table 5-39.

In the first column, we see the same person (Z) responding to five different topics (1, 2, 9, 10 and 11). The New ID column shows that every person x topic combination (e.g. Person Z x Topic 1; Person Z x Topic 2) is identified with a different ID reference. However, when the same person x topic combination is rated by different raters (in the case of Topic 9), then the New ID remains the same so that the resulting person x topic measures could

be adjusted for relative rater severity. The BK column shows the different BK groupings in the FACETS analysis whereas the BK Measure column shows the precise BK measures for person Z's BK of different topics. The Ability Measure column shows the results of ability estimates from a 4-facet MFRM analysis with (Examinee, Rater, Topic, Criteria) so the estimate remains the same for the person. In contrast, the Topic Ability Measure column shows the results from a 3-facet MFRM analysis (Examinee (person x topic), Rater and Criteria) illustrating how there might be fluctuations in the ability measures on the basis of the topic the person happened to respond to.

Table 5-39 Example of a Racked Data Set

PERSON ID	TOPIC	RATER	NEW ID	BK GROUPING	BK MEASURE	ABILITY MEASURE (10 TOPICS)	TOPIC-ABILITY MEASURE
Z	1	3	1	High BK	4.72	2.52	2.73
Z	2	4	2	High BK	5.458	2.52	2.70
Z	9	1	3	High BK	4.154	2.52	2.95
Z	9	2	3	High BK	4.154	2.52	2.97
Z	9	3	3	High BK	4.154	2.52	3.56
Z	9	4	3	High BK	4.154	2.52	2.85
Z	10	1	4	Medium BK	2.793	2.52	2.25
Z	11	2	5	Low BK	-3.565	2.52	2.02

Following the above procedure, the data for all the persons were racked and two separate data sets (with and without the TD criterion) were constructed. A 3-facet MFRM analysis was subsequently performed on the data. It is these resulting *person x topic* speaking measures which are used as the outcome variable in a multiple regression analysis and where the person x topic background knowledge measures are used as predictor variables. Additionally, and on the basis of the literature, a decision was made to include other predictor variables of general language proficiency (from the analyses of C-tests, discussed in the preliminary analysis chapter), task type and gender in the analysis. Before

reporting the results, an overview of multiple regression as a statistical technique is presented.

5.5.1 Multiple Regression Analysis: a Brief Introduction

Multiple regression analysis is a method of predicting an outcome variable from two or more predictor variables. It is suggested for predictor variables to be selected on the basis of ‘substantive theoretical importance’ (Field, 2010: 212). On the basis of the literature review, the variables of general language proficiency, background knowledge of topic, task type and gender were selected as the most important predictors of topic-based speaking performance in the current study. General language proficiency was identified as the most significant predictor.

5.5.2 Multiple Regression Analysis: Specifications

One of the considerations of the multiple regression analysis is sample size. There are different rules of thumb for a sample size necessary for a reliable regression line. Generally, 15 cases per predictor are recommended but Field (2010) warns that such rules of thumb run the risk of oversimplification of the issue and that the ‘sample size required will depend on the size of the effect that we’re trying to detect (...) and with how much power we want to detect these effects (Field, 2010:222). The estimation of R in a regression analysis is dependent on the number of predictors (k) and the sample size (N). Green (1991, cited in Field, 2010:222) recommends two rules of thumb for calculating sample size. One is used for estimating the fit of the regression model and the other is used for testing the predictor variables within the model (R^2 for the former and b -values for the latter). The following rules of thumb are recommended: **50+8k** for the overall model and **104+k** for the predictors.

To remind the reader, the data for the multiple regression analysis was raked, where every *person x topic* combination was treated as a distinct person, resulting in 809 cases (81 persons x 10 topics, excluding a corrupt file) available for fitting the regression line. There are four predictor variables (general language proficiency estimates, background knowledge estimates, task type and gender). The minimum sample size for the current study is calculated as follows.

$$50+8k= 50+8(4)=82 \text{ and } 104+k=104+4=108$$

Results show that the raked sample size (809) clearly exceeds the minimum sample size requirement of 108.

5.5.3 Method of Data Entry

The four predictor variables of general language proficiency, background knowledge of topic, task type and gender have all been identified in the literature as having potential influence on speaking performance with general language proficiency as the most significant predictor. A hierarchical method of data entry was therefore opted for. In this method the sequence of entering variables is driven by previous theoretical research and implemented by the researcher, as opposed to the computer program.

Two separate regression analyses were carried out with the same predictor variables but with two different outcome variables: a) person x topic speaking proficiency measures for all criteria and b) person x topic speaking proficiency measures for the IELTS criteria (without TD).

The predictor variables were entered in blocks. With the exception of task type, the three remaining predictor variables were either categorical with two categories (gender) or

interval, continuous data (proficiency and background knowledge). For task type with three categories, two dummy variables were defined and entered as one block in the model. The sequence of entry in the hierarchy was as follows:

- Block 1: General language proficiency estimates (from C-tests)
- Block 2: BK estimates
- Block 3: Task Type (A vs. B and A vs. C)
- Block 4: Gender

Different statistical options were selected in running the analyses in order to test for assumptions and fit of the model (e.g. collinearity diagnostics, Durbin-Watson test, Model fit, R squared change). Regression plots, histogram of standardised residuals and normal distribution of residuals were also specified in order to test for various other assumptions of the model (e.g. homoscedasticity and heteroscedasticity). A full discussion of what each of these terms mean is beyond the scope of this section. What is important is that when the assumptions are met ‘the model that we get for a sample can be accurately applied to the population of interest’ (Field, 2010:221). Results indicate that the assumptions of multiple regression were met (see Appendix T for these results).

5.5.4 Findings (FullA)

This section reports the findings from the multiple regression analysis. The outcome variable is the person x topic Rasch measures from a 3-Facet MFRM analysis with all five criteria.

Prior to evaluating the regression model, the correlation matrix was first examined in order to see whether there are very strong correlations ($r > 0.90$) between any two predictors which might give a rough idea of multicollinearity. None of the coefficients

were above 0.90 and therefore do not raise major concerns regarding multicollinearity at this stage.

The main output to consider is the regression model summary for each step of the hierarchy as predictor variables were entered (see Table 5-40).

Table 5-40 Regression Model Summary (FullA)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.776a	0.602	0.601	1.384275	0.602	1218.912	1	807	0.000	
2	.794b	0.631	0.630	1.333085	0.029	64.168	1	806	0.000	
3	.797c	0.636	0.634	1.326421	0.005	5.059	2	804	0.007	
4	.799d	0.639	0.636	1.32162	0.003	6.852	1	803	0.009	1.233
a. Predictors: (Constant), ProficiencyEst b. Predictors: (Constant), ProficiencyEst, BKEstimates c. Predictors: (Constant), ProficiencyEst, BKEstimates, TaskAversusTaskB, TaskAversusTaskC d. Predictors: (Constant), ProficiencyEst, BKEstimates, TaskAversusTaskB, TaskAversusTaskC, Gender e. Dependent Variable: 3-Facet 5 Criteria (FullA)										

The column labeled *R* shows the values of the multiple correlation coefficient between the predictor variables and the outcome variable. The value of R^2 in the next column is an indication of ‘how much of the variability in the outcome is accounted for by the predictors’ (Field, 2010:235). This value is 0.601 for Model 1, which means that general language proficiency accounts for 60.1% of the variation in spoken performance on different topics. Once all predictors have been entered into the model (Model 4) this percentage is increased to 63.6%, thus improving the first model by more than 3%. The adjusted R^2 column is an indication of the generalizability of the model and in ideal circumstances, the values for R^2 and adjusted R^2 should be very close, which is the case. For model 4, there is a reduction from 0.639 to 0.636 (0.003) which means that

approximately 0.3% less variance is accounted for by the model, if derived from the population, as opposed to the sample.

Field (2010:222) recommends using Stein's formula for cross-validating a regression model. Stein's equation is reproduced below:

$$R^2 = 1 - \left[\left(\frac{n-1}{n-k-1} \right) \left(\frac{n-2}{n-k-2} \right) \left(\frac{n+1}{n} \right) \right] (1 - R^2)$$

$$R^2 = 1 - \left[\left(\frac{809-1}{809-4-1} \right) \left(\frac{809-2}{809-4-2} \right) \left(\frac{809+1}{809} \right) \right] (1 - 0.639)$$

$$R^2 = 1 - 0.365 = 0.635$$

This value is very similar to the observed value of $R^2 = 0.639$, indicating that the cross-validity of this model is very good.

The change statistics are given in the next columns and indicate the significance of change in R^2 . The change in the F-ratio is reported for each block of the hierarchy so model 1 causes R^2 to change from 0 to 0.602, and this change in the amount of variance explained gives the F-ratio value of 1218.912, which is significant with a probability of less than 0.001. The addition of the new predictors has influenced R^2 to increase by 0.029 in Model 2 and 0.005 and 0.003 in models 3 and 4 respectively (the probability values are significant and less than 0.01 in each case). These change statistics are useful in designating the difference made to the model by adding additional predictors. The final statistic to consider is the Durbin-Watson statistic which 'informs us about whether the assumption of independent errors is tenable' (Field, 2010:236). Values less than 1 and

above 3 should raise concerns (*Ibid.*). The value in this analysis is 1.223, which suggests that the assumption has been met.

The next output to consider is the ANOVA table (Table 5-41) which tests whether or not the defined models are significantly better at predicting the outcome variable than using the ‘mean as the best guess’ (Field, 2010:236). For the initial model (1), the F-ratio is 1218.912, which is highly unlikely to have happened by chance ($p < .001$). The F-ratios for the remaining models are much smaller than the initial model but they are all significant at the 0.001 level, meaning that the new models (with the extra predictors) get increasingly better at the ability to predict speaking performance on different topics.

Table 5-41 ANOVA Output

ANOVA(e)						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2335.702	1	2335.702	1218.912	.000a
	Residual	1546.388	807	1.916		
	Total	3882.09	808			
2	Regression	2449.735	2	1224.868	689.245	.000b
	Residual	1432.354	806	1.777		
	Total	3882.09	808			
3	Regression	2467.537	4	616.884	350.623	.000c
	Residual	1414.552	804	1.759		
	Total	3882.09	808			
4	Regression	2479.505	5	495.901	283.911	.000d
	Residual	1402.584	803	1.747		
	Total	3882.09	808			
a. Predictors: (Constant), ProficiencyEst b. Predictors: (Constant), ProficiencyEst, BKEstimates c. Predictors: (Constant), ProficiencyEst, BKEstimates, TaskAversusTaskB, TaskAversusTaskC d. Predictors: (Constant), ProficiencyEst, BKEstimates, TaskAversusTaskB, TaskAversusTaskC, Gender e. Dependent Variable: 3-Facet 5 Criteria (FullA)						

What we are particularly interested in this analysis is the effect of BK. As illustrated, BK accounts for the biggest improvement in predicting speaking performance (3%) after

proficiency level. Moreover, on the basis of the ANOVA results, all predictors significantly contribute to the model. We can therefore define the model as follows (see the table of coefficients Table 5-42):

$$\text{Speaking performance on topics}_i = b_0 + b_1(\text{proficiency}_i) + b_2(\text{background knowledge}_i) + b_3(\text{task A vs. B}_i) + b_4(\text{task A vs. C}_i) + b_5(\text{gender}_i)$$

$$= -1.78 + (1.80 \text{ proficiency}_i) + (0.16 \text{ background knowledge}_i) + (-0.03 \text{ task A vs. B}_i) + (-0.32 \text{ task A vs C}_i) + (0.24 \text{ gender}_i)$$

The *b*-values are informative in terms of the relationship between speaking performance and individual predictors. The positive values of proficiency, background knowledge and gender suggest that as proficiency and background knowledge increase, speaking performance also increases. In terms of gender, the positive value suggests that girls perform better than boys (as girls were coded with 1 and boys with 0). Task type on the other hand has a negative relationship with speaking performance. This is as expected. An increase in the difficulty of task types is associated with a decrease in spoken performance scores. This relationship is not significant between task types A and B (there is only -0.03_i decrease, $p=0.811>0.05$); a finding which reflects MFRM results, as task types A and B were shown to exhibit similar difficulty levels. However, the relationship is significant between task types A and C ($p=0.003<0.01$).

The beta value for BK can be interpreted as follows: as BK values increase by one unit (one logit), spoken performance on topics increase by 0.16 logits. This interpretation is true when the effects of proficiency level, task type and gender are held constant. What becomes evident is that there needs to be a substantive increase or decrease in BK for it to have significant effect on spoken scores. This is in line with the results of the MFRM

analysis (with BK as a facet) where a significant main effect was shown for the BK conditions but even the maximum difference between the BK conditions fell below the minimum level of speaking ability necessary to move across score categories i.e. BK failed to have practical effect on speaking performance.

Another way of interpreting the results is to look at the standardised versions of the b -values (β), 'as they are not dependent on the units of measurement of the variables' (Field, 2010:239). Instead, they are reported in standard deviations i.e. 'the number of standard deviations that the outcome will change as a result of one standard deviation change in the predictor.

Proficiency (*standardised $\beta=0.762$*): This standardised Beta value indicates that as general language proficiency increases by one standard deviation (0.92), speaking performance on topics increase by 0.762 standard deviations. The standard deviation of speaking performance is 2.19 and this constitutes a change of 1.66 ($0.762 \times 2.19=1.66$). Therefore, for every increase of 0.92 logits in proficiency, there will be an increase of 1.66 logits in topic-based spoken performance.

Background Knowledge (*standardised $\beta=0.151$*): This value indicates that as background knowledge increases by one standard deviation (1.97), speaking performance on topic estimates increases by 0.151 standard deviations. The standard deviation for speaking performance is 2.19 and this implies a change of 0.33. This means that as BK increases by 1.97 logits, we can expect a change of 0.33 logits in speaking scores. This interpretation is only true if the effects of the other predictors are held constant.

The remaining predictor variables can be interpreted in the same manner.

Table 5-42 Table of Coefficients

Coefficients(a)					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	-1.67	0.05		-33.074	0.000
ProficiencyEst	1.837	0.053	0.776	34.913	0.000
2 (Constant)	-1.819	0.052		-34.941	0.000
ProficiencyEst	1.787	0.051	0.755	35.005	0.000
BKEstimates	0.191	0.024	0.173	8.01	0.000
3 (Constant)	-1.664	0.08		-20.729	0.000
ProficiencyEst	1.793	0.051	0.757	35.278	0.000
BKEstimates	0.168	0.025	0.151	6.695	0.000
TaskAversusTaskB	-0.03	0.128	-0.006	-0.238	0.812
TaskAversusTaskC	-0.327	0.108	-0.073	-3.027	0.003
4 (Constant)	-1.789	0.093		-19.19	0.000
ProficiencyEst	1.805	0.051	0.762	35.498	0.000
BKEstimates	0.168	0.025	0.151	6.729	0.000
TaskAversusTaskB	-0.031	0.127	-0.006	-0.24	0.811
TaskAversusTaskC	-0.326	0.108	-0.073	-3.033	0.003
Gender	0.244	0.093	0.056	2.618	0.009
a. Dependent Variable: 3-Facet 5 Criteria (FullA)					

5.5.5 Findings (IELTSA)

The same multiple regression analyses were repeated for the second data set (with the IELTS criteria and without TD). Here, only the model summary is presented. There were no violations of the model assumptions. Focus will be on the predictive ability of the background knowledge estimates.

The results of this multiple regression analysis (see Table 5-43) show that general language proficiency has remained the strongest predictor of topic-based spoken performance with an R^2 value of 0.601, accounting for 60.1% of variation. However, removing the TD criterion has decreased the predictive power of the remaining variables

so much so that Task Type and Gender are no longer significant predictors. Whilst still a significant predictor, the inclusion of background knowledge in the regression model has only improved the predictive power of the model by 0.8% to 60.9%.

Table 5-43 Regression Model Summary (IELTSA)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df 1	df2	Sig. F Change	
1	.776a	0.602	0.601	1.393	0.602	1219.6	1	807	0.000	
2	.781b	0.61	0.609	1.380	0.008	16.12	1	806	0.000	
3	.781c	0.61	0.608	1.380	0.001	0.769	2	804	0.464	
4	.781d	0.611	0.608	1.381	0	0.402	1	803	0.526	1.215
a. Predictors: (Constant), ProficiencyEst b. Predictors: (Constant), ProficiencyEst, BKEstimates c. Predictors: (Constant), ProficiencyEst, BKEstimates, TaskAversusTaskB, TaskAversusTaskC d. Predictors: (Constant), ProficiencyEst, BKEstimates, TaskAversusTaskB, TaskAversusTaskC, Gender e. Dependent Variable: 3Facet-4Criteria (IELTSA)										

These results are once again, in line with the MFRM results, demonstrating that the predictive power of background knowledge decreases once the TD criterion is removed from the analysis.

Taken together, the results of all analyses point to the same conclusion: topic and background knowledge of topic have a statistically significant effect on spoken performance. However, this effect is small and is not large enough to have a practical impact on scores. Background Knowledge of topic can predict 3% of variance in topic-based spoken performance. This predictive value falls to less than 1% when the TD criterion is removed from the analysis.

5.6 Interaction (Bias) Analyses

5.6.1 Interaction Analysis in MFRM: Introduction

In our analysis so far the facets of the assessment context have been examined independently in order to evaluate the extent to which each facet has an impact on measurement results. In other words we have looked at ‘main effect models’ (Eckes, 2009:32). It was outlined in the literature review chapter that interactions between different facets can also be a source of measurement error (McNamara, 1996) and that they need to be taken into account. In addition to main effects, MFRM allows for an investigation of ‘the interaction of as many of the original facets as the researcher wishes to in a single analysis (...) It is roughly equivalent to ANOVA’ (O’Sullivan and Rignall, 2007:454). In FACETS, interaction analysis is labeled *bias analysis* or differential facet analysis. For example, when the interaction analysis is related to items, the term differential item functioning (DIF) is used; when related to persons, it is called differential person functioning (DPF) and so and so forth. In reference to the conceptual-psychometric model (Eckes, 2009), we can see that a multitude of interaction analyses can be carried out between various features of different facets. In this respect, Eckes (2009:32) distinguishes between *exploratory* and *confirmatory* interaction analyses where the latter tests interactions on the basis of a previous hypothesis, theoretical rationale or research whereas the former is not based on a particular hypothesis and instead ‘aims at identifying systematic deviations from model expectations’.

A bias/interaction analysis is carried out by first measuring the different facet elements and rating scale structures, and subsequently anchoring (or fixing) those measurement

values. The expected value of observations are then compared against observed values; expressed in residuals. The residuals which correspond to the specified interaction terms (e.g. background knowledge by proficiency level) are subsequently summed. An interaction is observed if the sum does not equal to zero. In FACETS a *bias statistic* is reported which is an indication of the significance of the bias size: ‘this statistic provides a test of the hypothesis that there is no bias apart from measurement error. The bias statistic is approximately distributed as a *t* statistic’ (Eckes, 2009:32). Each bias term is associated with a measure, standard error, size and significance which allows ‘the effect of bias to be expressed in the same frame of reference as the element measures’ (Linacre, 2011a: 245).

5.6.2 Interaction Analysis: Background Knowledge x Language Proficiency

Previous research and theoretical rationales posited a possible interaction between background knowledge and general language proficiency. An interaction analysis was therefore carried out in FACETS (Linacre, 2011b) between these two specified facets in order to test this hypothesis and to answer the following research question:

viii. Does background knowledge of topics differentially affect performances of test takers from different proficiency levels?

The summary results of the bias analyses for FullA and IELTSA are presented in Table 5-44 and Table 5-45 respectively. It should be noted that background knowledge and general language proficiency measures were divided into three groups (Low, Medium and High) for these analyses, as FACETS only accepts integer number.

In the analyses specifications, a bias z-score value larger than |2| is flagged as showing significant bias. Results of the study indicate two significant bias terms (see the fourth

column in Table 5-44). If we look at the two columns on the far right we see that the first row is for persons belonging to a Low BK group but with a High Proficiency grouping. The positive bias measure of 0.11 can be interpreted in the following way: when background knowledge levels are low, higher proficiency participants are at an advantage compared to participants with low or medium proficiency levels. This can be inferred from the positive bias measure of 0.11. In contrast, when participants are both low proficiency and low background knowledge, they are at a disadvantage (bias measure = -0.1). We can therefore deduce that having low levels of background knowledge differentially influences persons from different the highest and lowest proficiency levels. While these bias terms are statistically significant, the size of the bias measures are very small and limited to 0.1 logits. It is therefore unlikely for these bias terms to have a significant impact on performance.

Table 5-44 Summary of Bias Analysis Results (BK x Proficiency), FullA

Raw Avg (obs-exp)	Bias Measure	Bias model S.E.	Bias Z- score	Bias infit MnSq	Bias Outfit MnSq	BK Group	Prof Level
0.05	0.11	0.05	2.01	1.3	1.6	Low BK	High Prof
-0.06	-0.10	0.05	-2.15	1	1	Low BK	Low Prof

Table 5-45 Summary of Bias Analysis Results (BK x Proficiency), IELTSA

Raw Avg (obs-exp)	Bias Measure	Bias model S.E.	Bias Z- score	Bias infit MnSq	Bias Outfit MnSq	BK Group	Prof Level
0.06	0.14	0.07	2.05	1.2	1.3	LowBK	HighProf

It should also be noted that the bias terms are not repeated for medium to high BK levels and for persons from medium proficiency levels. In other words having high levels of BK does not differentially affect persons from different proficiency levels. According to the results, it is only when BK is low that proficiency level plays a role.

For the IELTSA analysis, only one significant bias term is observed. Similar to the previous analysis, Table 5-45 illustrates that when persons of high proficiency have low background knowledge levels, they have a statistically significant advantage over persons with lower proficiency levels judging by the positive bias measure. The bias measure is nevertheless very small (0.14 logits) and is unlikely to have a substantial impact on measurement results.

In answering the research question posed at the beginning of the section, the results suggest that when all criteria are taken into account, there is a significant interaction background knowledge and language proficiency in the following cases: when background knowledge is low, persons of high proficiency levels are at an advantage compared to lower proficiency level examinees. Low background knowledge also puts persons at a low proficiency level at a disadvantage compared to examinees from higher ability levels. There was no evidence of an opposite trend, that is, high background knowledge was not shown to favour or bias against persons from different ability levels. It should be noted that despite reaching statistical significance, the bias measures were very small and therefore unlikely to have a large effect on performance.

When the same analysis was performed for the IELTS criteria (without the TD criterion), only one interaction term was found, which suggested a small yet significant advantage for the high proficiency examinees in relation to the low BK condition.

5.6.3 Differential Rater Functioning

In line with the foci of the study, a series of three additional interaction analyses were carried out in order to examine whether the raters in the study exhibited any systematic biases towards particular topics or whether they differentially rated the performances of persons from different levels of background knowledge of proficiency levels. The results of the analyses are used to answer the following research question:

- ix. Are there any significant interactions between raters and a) topics, b) test takers' background knowledge of topics and c) test takers' language proficiency levels?*

5.6.3.1 Interaction Analysis: Rater x Topic

The results of the rater x topic interaction analysis show three significant bias terms for FullA and IELTSA (see Table 5-46 and Table 5-47). In both analyses, Topic A.6 (Dancing; the easiest topic) is differentially rated by Raters 3 and 4. It appears that when test takers respond to this specific topic, they are likely to get a higher mark from Rater 4 than if they were rated by other raters. Conversely, the results for Rater 3 indicates that this rater, compared to the other raters, is likely to assign a lower score than the model expectations when rating performances on Topic A.6. A possible explanation for this finding is put forward by Lim (2009: 166) in that 'raters (may) internally adjust their rating behaviour depending on how difficult they perceive the (...) task to be'. This can also explain why Rater 3 exhibited a significant bias in the opposite direction for Topic B.2 (estimated the most difficult Task B topic). It is also possible that these finding reflect an interaction between these raters and *task types* perceived to be more or less difficult. A subsequent interaction analysis (rater x task type) showed only one bias term for Rater 3 (Bias size =-0.16; $z=-2.97$) for Type A tasks, confirming that this rater (the most lenient of

all four raters), exhibited more severity when rating Type A tasks (designed as the easiest tasks). The size of the bias measures in these analyses (between |0.43| and |0.80|) while relatively large, is unlikely to have a practical impact on scores as the maximum bias measure (0.80) in IELTSA falls below the minimum ability levels required to move across categories (see results of analyses in sections 5.3.7 and 5.3.8). Moreover, of all the possible rater x topic (4 raters x 18 topics) and rater x task type (4 raters x 3 task types) combinations, only few isolated cases of bias have been flagged up, which is a positive outcome. Nevertheless, the findings are suggestive of the different ways in which topics and tasks may interact with the severity/leniency of raters and generally illustrate some the complexities involved in an assessment context.

Table 5-46 Summary of Bias Analysis Results (Rater x Topic), FullA

Raw Avg (obs-exp)	Bias Measure	Bias model S.E.	Bias Z- score	Bias infit MnSq	Bias Outfit MnSq	Rater	Topic
0.26	0.50	0.15	3.28	1.4	2.0	R4	A.6
0.19	0.34	0.13	2.57	0.7	0.8	R3	B.2
-0.21	-0.43	0.15	-2.81	0.9	1.3	R3	A.6

Table 5-47 Summary of Bias Analysis Results (Rater x Topic), IELTSA

Raw Avg (obs-exp)	Bias Measure	Bias model S.E.	Bias Z- score	Bias infit MnSq	Bias Outfit MnSq	Rater	Topic
0.35	0.80	0.19	4.29	1.3	1.3	R4	A.6
0.20	0.45	0.18	2.53	0.9	0.9	R3	B.2
-0.22	-0.54	0.19	-2.88	0.6	0.8	R3	A.6

5.6.3.2 Interaction Analysis: Rater x Background Knowledge of Topic

The results of a bias/interaction analysis between the rater and background knowledge facets did not show any significant bias terms in FullA or IELTSA. This finding indicates that raters do not differentially rate the performance of test takers with different levels of background knowledge.

5.6.3.3 Interaction Analysis: Rater x General Language Proficiency

The largest bias z-scores are observed for the results of an interaction/bias analysis between raters and test takers' general language proficiency (see Table 5-48 and Table 5-49). Six bias terms are observed for both FullA and IELTSA with bias sizes ranging from $|0.19|$ to $|0.61|$. A close examination of the tables reveals a number of patterns. Firstly, unlike the previous analyses, all four raters exhibit bias, albeit in different ways. Secondly, raters exhibit bias towards the low and high proficiency groups but no significant bias terms are observed for the medium proficiency group.

The data in both tables show that Rater 3 (the most lenient rater) exercises more leniency when rating lower proficiency test takers, judging by the positive bias measure of 0.46. On the other hand, when rating test takers from the high proficiency group, this rater displays negative bias. In other words, as the test takers' proficiency levels increase, this rater exercises more severity. The opposite trend is observed for raters R2 and R4. These raters are systematically more severe towards low proficiency test takers but more lenient towards high proficiency test takers. Put differently, as the language proficiency of test takers increases, these two raters become more lenient. Rater 1 only displays a (negative) bias for the high proficiency test takers and is more severe towards this group.

The bias measures in this analyses are statistically significant yet unlikely to have a practical impact on performance, as the maximum bias measure (-0.61) generally falls below the average ability (approximately one logit or more) required to move across adjacent band levels (see results of Table 5-20 and Table 5-22).

Table 5-48 Summary of Bias Analysis Results (Rater x Proficiency), FullA

Raw Avg (obs-exp)	Bias Measure	Bias model S.E.	Bias Z- score	Bias infit MnSq	Bias Outfit MnSq	Rater	Prof Group
0.27	0.46	0.05	8.39	0.7	0.7	R3	LowProf
0.22	0.44	0.06	7.33	1.5	1.7	R4	HighProf
0.12	0.24	0.06	4.03	0.8	0.8	R2	HighProf
-0.09	-0.19	0.06	-3.31	1.3	1.3	R1	HighProf
-0.12	-0.22	0.05	-4.11	0.9	0.9	R4	LowProf
-0.2	-0.38	0.06	-6.31	0.9	0.9	R2	LowProf

Table 5-49 Summary of Bias Analysis Results (Rater x Proficiency), IELTSA

Raw Avg (obs-exp)	Bias Measure	Bias model S.E.	Bias Z- score	Bias infit MnSq	Bias Outfit MnSq	Rater	Prof Group
0.23	0.58	0.07	7.78	1.7	1.8	R4	HighProf
0.22	0.44	0.07	6.62	0.7	0.7	R3	LowProf
-0.11	-0.23	0.06	-3.6	0.9	0.9	R4	LowProf
-0.1	-0.24	0.07	-3.47	1.3	1.3	R1	HighProf
-0.18	-0.38	0.07	-5.39	0.9	0.9	R2	LowProf
-0.22	-0.61	0.08	-7.79	0.6	0.6	R3	HighProf

In answering the research question posed at the beginning of the section, the findings from the interaction analyses illustrate the myriad ways in which interactions between different elements of the assessment context can influence measurement. For example, isolated cases of interactions were observed between specific raters and specific topics. No bias terms were observed between raters and test takers' background knowledge of topics. The most striking trend was detected in the two distinct ways in which raters' severity/leniency

interacted with test takers' proficiency levels: raters exhibiting more leniency for the low proficiency group but more severity for the high proficiency group vs. raters exhibiting more severity for the low proficiency group but more leniency for the high proficiency group.

While all the observed bias terms were statistically significant, they were not large enough to have a practical impact on performance scores.

5.7 Test-Taker Perspectives

The quantitative analyses of speaking performance scores have generally suggested that topic and background knowledge of topic can have an effect on performance. However, despite exhibiting statistical significance, the effect has failed to reach practical significance. In other words, it is unlikely for test takers to achieve higher or lower scores as a consequence of the choice of topic they are assigned or the amount of background knowledge they bring to the topics.

This section moves away from the analyses of observed scores and instead focuses on test-takers' perspectives of the role of topic and background knowledge of topic in relation to their final scores. More specifically, the findings reported in this section aim to answer the following research question:

- x. What are the test takers' perceptions of the influence of topic and background knowledge of topic on performance?*

To remind the reader, four statements, designed to elicit test takers' perceptions of topic effects on scores, were included in section II the background knowledge questionnaire (discussed in 3.3.2.2) and are reproduced below.

1= Strongly disagree, 2= Disagree, 3= Undecided, 4= Agree, 5=Strongly Agree

1. You think that the choice of topics might affect your final score.
 2. You think that having more ideas about a topic might affect your final score.
 3. You think that there is an element of 'luck' involved in the choice of topics.
 4. You think that the choice of topic is not important if your English is good enough.
-

This section reports the results of the analyses of test taker responses to these statements using descriptive statistics. The distribution of responses to each question is visually presented in Figure 5-10.

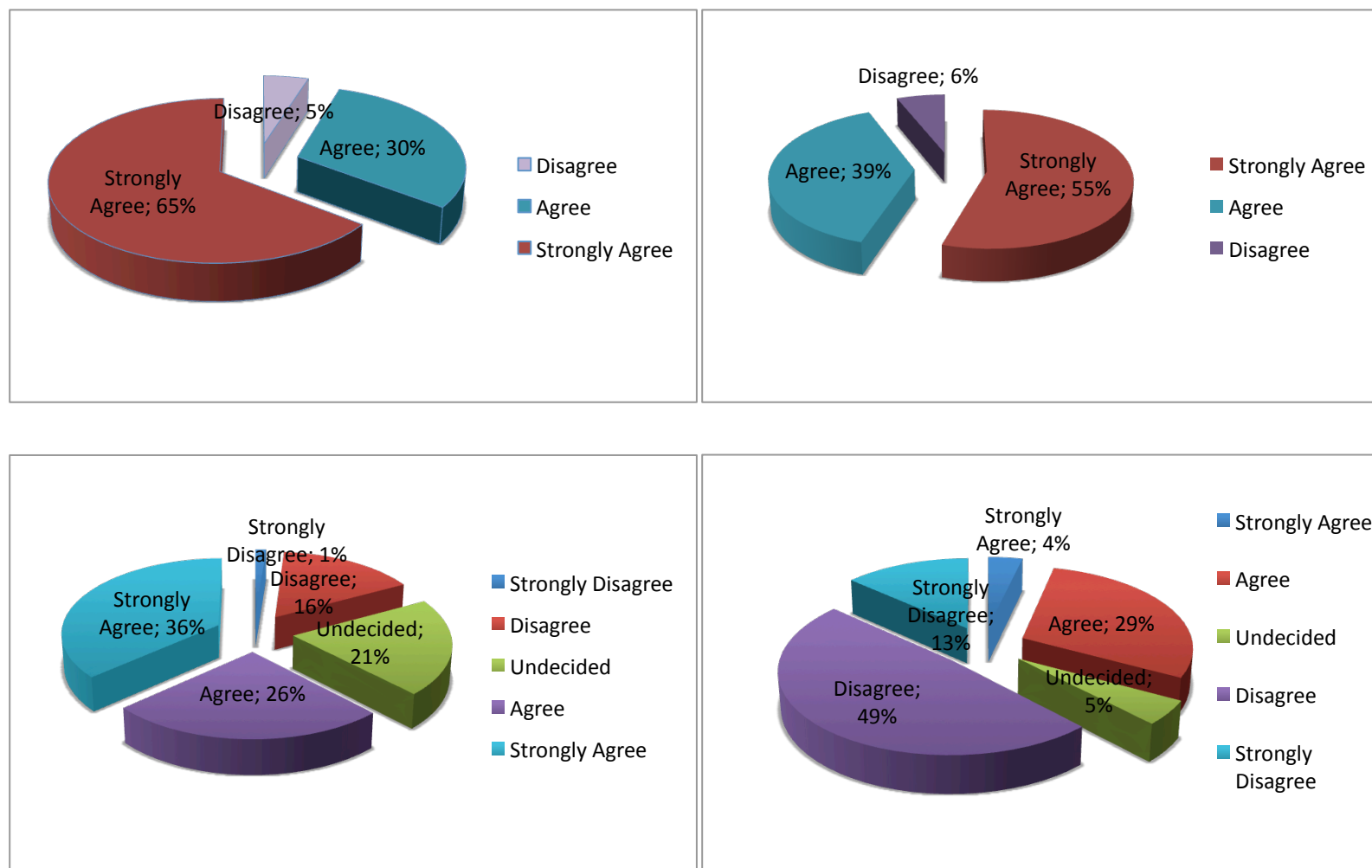
Results for statements 1 and 2 indicate that a striking majority of respondents (approximately 95%) believe that choice of topics and having more ideas about a topic might affect their final score. The *strongly agree* option was selected by 65% and 55% of the participants (the majority) for statements 1 and 2 respectively. Only 5-6% of respondents disagreed with these statements with no respondents opting for the *undecided* option. The test takers' strong attitudes towards the impact of topic and background knowledge of topic on their performance were also evident in additional comments by some of the test takers in response to these statements; comments such as 'I passionately agree' or 'I couldn't agree more' or '100% agree'.

The third statement aimed to elicit the extent to which test takers associate the topics assigned to them in language proficiency interviews with luck of the draw. Once again, the majority of respondents (62%) thought that there is an element of 'luck' involved in the choice of topics. About 21% were undecided and 17% disagreed with the statement.

The aim of the fourth statement was to examine whether there is a change in the pattern of test-taker responses in relation to the effects of topic on performance scores, once they are

asked to consider the role of language proficiency. Results suggest a shift in the pattern of responses from statement 1 to statement 4. While the majority of respondents still disagree with the negatively worded statement ‘the choice of topic is not important if my English is good enough’, this percentage had drastically dropped from 95% to 62%. Moreover, an increase is observed in the percentage of respondents who felt that choice of topic is no longer an issue at high levels of proficiency (29% agree and 4% strongly agree).

The findings from the analyses of test-taker questionnaires reveal an inconsistency between the *perceived* impact of topics and background knowledge of topics on scores on the basis of the test taker questionnaire results and the *observed* impact of these variables on scores, on the basis of the MFRM results. While test takers believe topic and having ideas about a topic to have an effect on their final scores, the MFRM results have consistently shown otherwise. These contrasting findings are discussed in more depth in Chapter 7.



Statement 1: You think that the choice of topics might affect your final score (top left)
Statement 2: You think that having more ideas about a topic might affect your final score (top right)
Statement 3: You think that there is an element of 'luck' involved in the choice of topics (bottom left)
Statement 4: You think that the choice of topic is not important if your English is good enough (bottom right)

Figure 5-10 Distribution of Questionnaire Responses

5.8 Summary of Findings

This chapter reported on the quantitative results of the study. The MFRM analyses allowed for a systematic examination of the various facets of the assessment context.

Findings from the examinee facet suggested a wide distribution of statistically distinct ability levels in the participants which implies that, on the one hand, the study was successful in selecting participants from a variety of oral proficiency levels and, on the other, that the speaking tasks in the study were successful in distinguishing between test takers from different abilities.

The rater facet results showed that the four raters in the study exercised severity levels which were significantly different from one another. Nevertheless, raters in the study exhibited high levels of consistency in marking and did not display systematic rater effects such as extremism, central tendency or individual halo effects. It was also suggested that a halo effect common to all raters might exert an influence on awarded scores due to the common structure of the rating scales.

The criterion facet results suggested that the five criteria in the study exhibited statistically distinct levels of difficulty and contributed in distinct ways to the separation of test takers into different ability levels. An exploratory examination of the Topic Development criterion suggested that the criterion appeared to function as intended in isolating the effects of topic-related factors. Attention was drawn to the criterion's markedly lower difficulty measure and relatively higher misfit statistics compared to the other criteria as potential evidence of the criterion's contribution to multidimensionality. However, the criterion's fit statistics, which fell within the acceptable range, implied that the evidence

was not strong enough to raise serious concerns and that the criterion generally worked well with the other criteria in distinguishing persons from different speaking ability levels.

An examination of the rating scale structures and categories demonstrated that the categories within the scale were generally functioning as intended.

Findings from the MFRM results of the topic and background knowledge facets were drawn upon to answer the majority of the study's research questions. At the test level, the topics in the study exhibited difficulty measures which were statistically distinct. This was as expected in light of the fact that the tests included task types designed to increase in difficulty. However, the sequencing of topics in difficulty measures did not match the expected progression in task type difficulty.

On the basis of topic difficulty and for comparative reasons, two forms of the test were constructed consisting of the easiest vs. the most difficult topics. Results suggested that even when the two speaking test forms included the easiest versus the most difficult topics, the performance measures were unlikely to be influenced by differences in topic difficulties as, on average, the minimum speaking ability required to move across adjacent band levels in different criteria consistently exceeded the maximum difference between the two test forms.

Topics were next examined at the task level. Results indicated that topics within each task type can be generally divided into a minimum of two statistically distinct difficulty levels and therefore cannot be considered parallel. However, when differences in topic difficulty measures were examined in relation to the speaking ability necessary to move across adjacent band scores, it was illustrated that these differences were not large enough to translate into differences in scores. In other words, the effects of topic were found not

have practical significance in terms of test takers achieving a higher or lower band scores. Differences in topics within each task type were also examined in relation to the functions they elicit. Findings showed both similarities and differences in the range of functions that different topics elicit. The results of the analyses of functions across the topics and task types also revealed the information-oriented nature of the speaking tasks as evidenced in the dominance of the informational functions relative to the interactional functions.

When background knowledge of topics was modeled as an additional facet of the assessment context, findings suggested a significant impact of background knowledge on performance. Differences in background knowledge of topics were shown to pose statistically distinct levels of challenge for test takers. However, similar to the results of the topic facet, background knowledge failed to exhibit practical significance in terms of an impact on performance scores. The maximum difference between the BK conditions was consistently lower than the minimum speaking ability required to move across adjacent bands for the different criteria. The only exception was observed for the TD rating scale in which BK levels could potentially exert an influence on achieving higher or lower scores at specific levels of the scale. The results of a multiple regression analysis showed that BK generally accounts for approximately 1-3% of the variance in test takers' performances on different topics and that a substantive increase in BK can result in a small change in speaking scores thus substantiating the results of the MFRM analyses.

Background knowledge was also shown to have a significant interaction with general language proficiency. In cases where background knowledge of a topic is low, higher proficiency test takers were found to be at an advantage compared to lower proficiency

examinees. High levels of background knowledge on the other hand were not shown to favour or bias against persons from different ability levels.

Significant interactions were also found between test takers' language proficiency and raters' harshness and leniency with raters either exhibiting more leniency for the low proficiency group but more severity for the high proficiency group or vice versa. These interactions, amongst others, illustrated the complexities involved in an assessment context and the possible sources of measurement error which can in turn affect the interpretations made on the basis of scores.

In contrast to the results of the quantitative analyses of score data, which consistently suggested a lack of a practical impact of topic and BK on scores, the results of test taker questionnaire response data showed that test takers attribute a significant role to both these variables in effecting their scores.

The results of this chapter have shown that topic and different levels of background knowledge of topic can have a statistically significant effect on scores. However, this statistical significance failed to translate into practical significance, as the size of these effects were shown to be very small compared to the speaking ability levels required to receive higher or lower band scores across the different criteria. Taken together, these results provide strong support for the topic validity of the speaking test under examination, as scores on the speaking tasks appear to predominantly reflect the underlying speaking ability construct that the test is designed to measure. The results reported in this chapter will be brought together with the study's qualitative findings before drawing conclusions in the final chapter.

Chapter 6 Findings (Qualitative)

This chapter presents the results of the study's qualitative findings in answering the study's final research question formulated as follows:

- xi. What themes and patterns emerge from a qualitative inquiry into a) rater perspectives and b) the content of test taker speaking performances, in relation to topic validity aspects of the speaking test under examination?*

Qualitative research is often contrasted with quantitative research where the strength of the former is in shedding light on aspects of the phenomenon under investigation which cannot necessarily be captured in statistical procedures (Creswell, 2003). The qualitative analysis of data was carried out with the following objectives in mind: a) generating themes and evidence that can help build a topic validity argument for the speaking test under examination and answer the study's overarching research question and b) forming 'explanatory patterns' (Ellis and Barkhuizen, 2005:257) which can provide an insight into the quantitative results of the study. The findings from the two strands of analyses are then converged in addressing the study's overarching research question (Creswell, 2003; Creswell and Plano-Clark, 2007) in the next chapter.

6.1 Qualitative Data and Analyses

6.1.1 Data Sources and Sampling

Two sources of qualitative data are explored in this study. The first source is the raters' interview transcript data. As explained in the methodology chapter, upon completion of the rating process, the raters in the study were interviewed about their rating experiences and their views on the influence of topic on speaking, having rated approximately 200 topic-based performances. As the data consisted of short interviews with the four raters, no sampling step was necessary and all transcripts were used for further analyses.

The second source is the test takers' spoken data. Within the field of applied linguistics, second language acquisition and language assessment, spoken data are predominantly used in two distinct ways as identified by Ellis and Barkhuizen (2005:359): on the one hand, they can be used 'as a source of data' for investigating what learners or candidates 'know and can do with an L2', and on the other hand, they can serve 'as a source of information' for investigating factors related to L2 learning or L2 performance in assessment contexts. In this study, it is the latter (content of spoken data) which is the focus of analysis. In the previous chapter, the participants' spoken performances were transformed into scores using the rating scales. However, the content of the same spoken performances can serve as a possible source of rich data, in the form of meanings expressed by candidates when talking about different topics, which can be explored using qualitative techniques.

Given the large number of spoken performances and one of the main concerns of the study i.e. the influence of *lack* of background knowledge, a sampling decision was made to

select those performances where topic-related problems in terms lack of background knowledge was most likely to have been in effect. Two sources of data – on the basis of the quantitative results of the study – were used as criteria for sample selection:

1. Background knowledge questionnaires

Performances on topics for which participants' self-reports of background knowledge of topic fell into the 'low background knowledge' category (on the basis of Rasch-based BK estimates) were identified and extracted. There were a total of 96 performances. For approximately 70% of the participants, there was at least one topic for which the level of BK was low.

2. Table of Unexpected Responses

The findings chapter suggested that some of the performances flagged in the table of unexpected responses for exhibiting high residual statistics were those in which the observed score on the topic development criterion was significantly lower than the expected score (see section 5.3.10 and Appendix S). Given that the topic development criterion is the criterion most likely to absorb topic-related effects, it was hypothesised that the large deviations between the observations and expectations of the model as well as the direction of the deviations may be attributable to a background knowledge effect. On this basis, 36 performances identified in the table were subsequently selected.

When the two data sets were crosschecked, an overlap of eight performances was found between them bringing the total to 121 performances out of 809 (15%). The data were subsequently transcribed for analyses.

6.1.2 Qualitative Analyses

The study's qualitative data sets (rater interview transcripts and speaking test transcripts) were analysed following a series of qualitative steps (Strauss and Corbin, 1998; Ellis and Barkhuizen, 2005). In the first instance, the transcripts were browsed through and a careful inspection of the data was carried out in order to get an overall impression of the data. The next step of the analysis consisted of a recursive process of note taking, sorting similar materials, labeling codes and modifying and re-assigning of codes, all of which aimed to capture similar meanings relevant to the foci of the study in the contents expressed in the transcripts. The coding process, which is an abstraction of the content of qualitative data to higher order concepts (Ellis and Barkhuizen, 2005), helped reduce the large sums of text to smaller, more manageable codes. Patterns and themes were then identified by establishing links through similarities in the coded data. These common themes (and their interpretation) constitute the qualitative findings of the study. However, in line with Goetz and Le Compte (1984), who invite researchers to present instances of data that may contradict the identified common themes, counter examples are also reported, where relevant. These examples are borrowed from data transcripts where participants' background knowledge of topics fell in the 'high' category. This type of evidence serves to 'establish the parameters or distribution of a construct' (Goetz and Le Compte, 1984:175).

It was explained earlier that the rater and speaking test transcripts were separately analysed. However, an examination of their respective findings revealed themes which were either common across the analyses or which provided complementary perspectives

for data interpretation. For these reasons, the study's findings from the two sources of qualitative data were synthesized and are presented together in the next sections.

6.2 Findings

A total of nine themes emerged from the qualitative analyses of data. In the following sections, each of these themes is expanded upon and illustrated with representative extracts from the transcripts.

- Dealing with topic unfamiliarity: topic abandonment and test-taker strategies
- The role of general language proficiency
- Topic unfamiliarity and negative affective influence
- Topic unfamiliarity and impact on performance and ratings
- Topic unfamiliarity and impact on rater decision-making
- The mediating role of task format
- Problematic topics and questions: the issue of local validity
- Topic interactions with characteristics of the individual test taker
- Test constraints and lack of a support mechanism

6.2.1 Dealing with Topic Unfamiliarity: Topic Abandonment and Test-Taker Strategies

The analyses of the content of test taker performances suggest that when faced with an unfamiliar topic and in replying to questions for which they do not necessarily have an answer, test takers tend to explicitly signal their lack of background knowledge. This is sometimes followed by complete topic abandonment where test takers fail to elaborate on a response. Examples of questions and candidate responses are presented below:

Q: Where can people in your country get information about genetic research?

TT09⁴⁹: *'I don't know that where we should go and ask about anything, about genetics'*

TT017: *'Uh I don't know, I don't know if there is any genetic research in Iran, I have not heard of that'*

TT032: *'Genetic research? I don't actually know'*

TT060: *'Where can get genetic research. (.) I don't know'*

Q: Have there been any changes in the number of jobs available in fishing and water transport industries do you think?

TT044: *'Mm I really don't have any information about this, I don't know'*

Q: Can you tell me about any traditional dancing in your country?

TT069: *'I have uh not enough knowledge about traditional dancing in Iran' 12*

In reviewing these responses, we can also rule out the possibility of the test takers' listening comprehension contributing to topic abandonment as the test taker either checks comprehension by repeating the question or by rephrasing the question in the response, as exemplified in responses by test takers TT017 and TT060 above.

Instead of disengaging from a topic or providing minimal answers, some test takers adopted different strategies in dealing with an unfamiliar topic, for example, by providing explanations for their lack of topic-related background knowledge, waffling, speculating or going off on a tangent. An illustrative example of a test taker exercising a variety of these strategies is presented below:

TT001: *'That's a very strange question, like I don't know, I'd say I have no clue about genetic research, and I have no idea about it, I don't even know if I like it or not, so I don't know what my people would think, I mean if you ask my people what they think about money, or what they think about freedom, or newspapers or television or soap operas they definitely have an answer, but if you, I can imagine those people who ask this question from and they will look at you baffled, like they have no idea what you are talking about, because genetic research, I think, belongs to genetic researchers, not to general people, and I don't think people buy newspapers every day to see the new advances in genetic research.'*

⁴⁹ Test takers are identified by TT and a code number.

In this extract, the test taker repeatedly emphasises on his lack of background knowledge. By way of explanation he then refers to the unfamiliarity of topic for his ‘people’ i.e. Iranians in general. Disregarding the original question, he then shifts the topic by elaborating on what people would be able to generally talk about before drawing attention to the irrelevance of the topic for those without specialist knowledge.

These findings closely reflect rater comments on the different ways in which test takers deployed strategies in dealing with unfamiliar topics or problematic questions.

Rater Extract 1

When people couldn't really talk about the question they would talk about something tangential just to be talking.

Rater Extract 2

Some would say 'I don't know anything about that' or speculate saying 'I guess', 'maybe', 'this could have something' you know something like that.

Rater Extract 3

I mean sometimes the problems with the content ended up with people kind of waffling about something.

6.2.2 The Role of General Language Proficiency

The raters in the study drew on the familiar theme of general language proficiency as a possible explanation for the differences between the ways test takers approached topic-related difficulties (e.g. topic abandonment or adopting strategies). It was suggested that lack of background knowledge might differentially affect test takers from high and low proficiency levels:

Rater Extract 4

It seemed that the higher performers kind of found it funny when they were given stuff they couldn't talk about and they were upfront about that and maybe for low performers they weren't able to figure out if what was going on was their English or whether they were given a question they couldn't talk about whereas the higher performers would be 'oh I don't know anything about that that's ridiculous or whatever whereas low performers were like oh maybe I should know something about that or maybe I haven't understood the question. It did seem like a sort of stranger topic was more problematic for a lower performer because maybe they weren't as aware that they were being asked something strange they were just afraid all of a sudden. A certain feeling of what am I going to do about this whereas it seemed like a high performer was more able to explicitly say that they didn't know anything about it, explain why and then maybe talk about something related enough you know like use a strategy.

These findings also lend support to the quantitative findings of the study. The results of a bias interaction between test takers' general language proficiency and background knowledge of topics showed that low levels of background disadvantaged test takers from low proficiency levels compared to those from higher language abilities. The insight from the qualitative findings suggests that unlike higher proficiency test takers, lower proficiency individuals may not be able to deploy strategies that allow them to deal with problematic topics, thus precipitating anxiety and increasing the likelihood of topic abandonment.

6.2.3 Topic Unfamiliarity and Negative Affective Influence

From the perspective of the raters, topic unfamiliarity was seen to have a negative affective influence on test takers, particularly those with lower proficiency levels (see previous section). Rater comments suggested that when encountering an unfamiliar topic, some individuals appeared to 'be caught off-guard', feel 'baffled', 'surprised' or at times, 'scared'.

Rater Extract 5

I think it [topic] can [have an effect] especially for the low performers and it can be quite scary to be given a topic that seems really remote.

6.2.4 Topic Unfamiliarity and Impact on Performance and Ratings

The adverse impact of topic unfamiliarity on features of spoken performance on the one hand, and on scores awarded to test takers on the other were noted by the raters. In relating performances to the rating scale criteria and the functions checklist, the influence of topic unfamiliarity was often associated with an increase in pauses and hesitations (FC scale), a decrease in lexical diversity (LR scale) and grammatical complexity (GA), poor topic development as well as a drop in the range of observed functions. Pronunciation was pointed out as the scale least likely to be negatively affected by lack of background knowledge while a negative ‘knock-off effect’ on ratings for the remaining scales was predicted. Illustrative rater comments are reproduced below.

Rater Extract 6

As an examiner and as a rater though, I could see how that [topic unfamiliarity] could have a negative knock-off effect on the overall rating because it did (.) when there was a pause it did lend itself to disfluencies, more pausing, kind of struggling to say something and you know topic development and the other descriptors then it was harder to get higher rating.

Rater Extract 7

I think the more random questions I think it was kind of in terms of them hesitating and backtracking and stuff and I'd probably be exactly the same and English is my first language.

Rater Extract 8

It's sort of not enough to evaluate their grammar and fluency because they can't really provide a complex...they can't use the subjunctive a lot to say why they don't anything about this. It's often going to be in very simple grammar, when they know nothing about it.

Rater Extract 9

Well there would be almost nothing in the observation checklist, there would be like asking for topic clarification and then comment on its difficulty and that was it.

Rater Extract 10

Often times, if they were high performers they would still receive a good mark for pronunciation because they would do those things with immaculate pronunciation but then when it came to something like lexical resource, like mmm they used no grammar, no vocabulary related to accuracy would also be really low because they usually used you know present simple or something like that u:m so they would do well with pronunciation but then it would look really bad for their like topic development and their lexical resource and those would look really bad.

6.2.5 Topic Unfamiliarity and Impact on Rater Decision-Making

The analyses of test taker performances showed that when faced with a problematic topic, some test takers abandoned the topic and failed to elaborate on a response while others opted for different strategies. The analyses of rater comments illustrate that the effects of topic unfamiliarity are not limited to the test takers but that they can pose a number of challenges on the raters in their decision-making process. For example, when a topic or question failed to generate enough samples of speech from test takers, raters remarked on an additional cognitive burden on their rating process where they had to ‘guess’ or ‘take a stab’ at what test takers can ‘potentially’ do with the language. However, it is not merely the length of the response that is a problem but the dilemma of rating responses which were considered by raters to be ‘short yet perfectly adequate and legitimate’. Raters’ attitudes towards strategy use were also mixed with one rater ‘rooting’ for those respondents who used strategies to circumvent topic-related problems while others expressed negative views of strategies such as waffling and going off on a tangent. In this regard, topic development was viewed as a helpful criterion in making scoring decisions for such responses.

Rater Extract 11

I was almost rooting for those respondents who turned around and said oh well I’m not really interested in that I have no idea...I can guess (.) and I thought well fair enough.

Rater Extract 12

There were times um there were ones very fluent not necessarily accurate and gave a very very long answer and at the end when you think back to the question they were asked it really bears no relation whatsoever to the question they were asked so I think it's a good idea to have a topic development criterion.

The different ways in which raters view and mark short and/or irrelevant responses may in fact be a reflection of more substantial differences in the ways they regard the role of topics from the perspective of test purpose:

Rater Extract 13

Even if they [test takers/respondents] are not on topic, at least they're talking and um or I know from experience it's difficult to keep to the point and you talk about something else and you kind of wonder with that. I mean the tasks are supposed to be there to generate talk and therefore talk must be surely the most important thing and if they go off on a tangent I think it might show up in the topic development score that they haven't actually addressed the task properly but you get reasonable marks in the rest of it because they spoke and you can judge their pronunciation and grammar and so on.

Rater Extract 14

As a rater or an examiner I think it's a good idea you're rating what language they use and how they use it but if they're not answering the question so what's the point? It's kind of like a conversation and if you don't answer the question then how much coherence is there really?

The above extracts suggest that some raters may view the principal function of topics as generating speech so that the linguistic aspects of a performance could be evaluated regardless of test takers' communicative success in addressing the topic. In contrast, others may regard the communicative purpose of the test as equally important – if not more so – than the linguistic aspect.

6.2.6 The Mediating Role of Task Format

The findings discussed so far have highlighted how topic unfamiliarity and lack of background knowledge can pose challenges for test takers and raters alike. Moreover, the raters attributed an important role to topic unfamiliarity in negatively influencing test taker

performances and scores. However, these results run counter to the quantitative findings of the study, as topic and background knowledge of topics were not shown to have a practical effect on speaking scores. An excellent explanation for these seemingly contradictory findings is found in an insightful remark by one of the raters who identified the multi-question format of tasks and the use of communication strategies as factors that can minimise the (negative) influence of lack of background knowledge on scores:

Rater Extract 15

It's clear to me that the test can get away with questions like 'where can you get information about genetic research in your country?' because even if candidates plead ignorance and know absolutely nothing about it, there are follow-up questions that supposedly will bail them out, and then they can, in theory, use their proficiency combined with strategic competence to pull it all off. However what I've noticed during the rating process is that that initial ignorance has a negative effect on both the candidate and the rater. There is no question in my mind that when a candidate starts off with major disfluencies that it has a knock-on negative effect on the rest of the rating. It is, after all, largely based on the impression of a small sample of speech. Conversely, and even worse, even though as I mentioned a candidate should in theory be able to deploy metacognitive mechanisms to recover when the second question comes, that does not happen, or doesn't happen well, because the candidate's working memory has gone into overload from the ignorance and stress of the first question.

This observation is largely supported in the qualitative analyses of the content of performances where test takers face problems in answering a specific question within a topic sequence. Put differently, the multi-question format of the tasks ensures that even if one or two questions fail to generate speech, the test takers are given enough opportunities to respond to at least some of the questions. This design therefore increases the likelihood of generating sufficient samples of speech on the linguistic criteria thus moderating the negative influence of lack of background knowledge on scores. What can also explain the discrepancies between the *perceived* influence of topic-related factors on scores (on the basis of test taker questionnaire results and rater interviews) and the lack of an *observed* impact on scores (on the basis of MFRM results) is that an unfamiliar question or topic

leaves such a strong, negative and lasting impression on test takers and raters alike that a direct effect on scores is automatically assumed.

Moving away from the impact of topic-related problems, the next section focuses on specific topics and questions which were found to be problematic in an attempt to identify the sources of the problem.

6.2.7 Problematic Topics and Questions: the Issue of Local Validity

The analyses of the test taker performances showed that the majority of the problematic questions or topics were related to the three topics of ‘Festivals’ (A.3), ‘Genetic Research’ (C.8) which required test takers to draw on previous experience or knowledge which they, as a group, did not necessarily have available on account of their Iranian background and other cultural and religious factors. The following extracts illustrate that many test takers found these topics irrelevant to their local context:

Q: Tell me about the most important festival in your country.

TT050: *‘Well uh as you know uh I am Iranian and in Iran actually we don’t have any festivals’*

TT078: *‘Mm we don’t have any special festivals, maybe uh 22 Bahman.’*

TT018: *‘Um we don’t have that much festival in Iran, and I think, um it would be better if we have some more festivals, more fun festivals actually. (18)’*

TT011: *‘Oh, let’s think of a festival. [LAUGHS]. Sorry, I really don’t remember.’*

TT030: *‘Actually I think we don’t have any festivals in our country. We have, I don’t think so.’*

TT033: *‘I think that would be uh [00.11] because it’s not so much uh the people in our country, and the government, are not so open-minded for different kinds of festivals and stuff in this country, I think so. At this I am interested in, for example, a lot of artistic festivals but that don’t exist so far fully in Iran.’*

TT040: *‘Festival? I don’t know, because most of the festivals in Iran are religious, and I’m not interested in that [LAUGHS] and I really don’t know.’*

Some test takers name festivals but they appear to be uncertain of whether their choices can be counted as festivals and then provide further explanations for why the topic may not be popular – explanations which are mostly tied to socio-cultural and/or religious reasons:

Q: Tell me about the most important festival in your country.

TT003: *The most (.) festival is film and I don't remember anything else, the films and somehow the football games somehow.*

TT004: *'I don't know is a festival or not (.) is a festival that last uh Tuesday after a year and everybody burned a mm (1.0) wooden and they jump on the wood and uh (.) they believe that when they jump uh on the fire uh they leave all of sicknesses and diseases and they begin to relive very fresh life in the New Year. It's one of our festivals.'*

TT007: *[LAUGHS] It's called [0.15] 22 Bahman [LAUGHS]. Maybe not a festival but festival that government in fact celebrates the victory of the Islamic group, of the people than the dictators Shah and the entrance of leader. You know, sometimes the festival was very [unintelligible word] but these days became a kind of festival, just festival name it and people doesn't mention it and even don't care about it, you see.'*

TT010: *[LAUGHS] You know we used to have many festivals but nowadays many of them are not held anymore, and many of them are not known anymore, but I will talk about a festival that is called [Chaharshanbeh Soori] is made of two words [test taker then elaborates] and this festival is held once a year, and nowadays it's forbidden in Iran, because Islam says that we should believe in only one God, and the old religion believed in God and the Divine, which had equal power.'*

Q: What special food and activities are there in this festival?

TT033: *'Uh I don't know about the food and stuff, but I know about [00.46] any time of year you go (test taker elaborates here) for example they do a lot of great food exhibitions that was so nice, but they are not that big to be compared to festivals or counted as festivals.'*

Q: Do you think that festivals are important for a country?

TT050: *'My opinion it depends on culture of any country, uh for example for Islamic countries maybe it is not very important for people, and for some Western country it will be more attractive, besides they are looking for some more fun and you cannot find such a culture in Islamic countries, that's it.'*

The raters in the study also commented on the irrelevance of the Festivals topic for the Iranian speakers in the sample. They also questioned the 'fairness' and 'local validity of

this topic, in particular, in comparison to other contexts where festivals may be more popular:

Rater Extract 16

'Oh yeah there was that festival oh yeah but she didn't really know much about it so she couldn't really talk about so I thought well what was the motivation behind that question? And you know did the person who decided that was a good idea, were they working in a country where festivals were prevalent?'

Rater Extract 17

'I think also like the Iranian thing may be different as well like I think it was the festivals one (.) um a lot of them said well we don't really have any festivals and I just thought that was really unfair because if it was like Chinese speakers or like Spanish speakers or somewhere where there are tons of festivals that would be an unfair advantage.'

A similar trend was observed for the 'Genetic Research' topic, which, once again, required test takers to draw on knowledge or information that they did not have available. Most test takers found the topic and the questions strange and puzzling and referred to its unpopularity and lack of topicality amongst Iranians and in the news and media.

Q: Where do people in your country find information about genetic research?

TT009: *'Um I think in our country genetics is actually it's not very um not very um popular or not very um uh we don't have any centre of genetics [00.30] centre of genetics in our country to go there and ask a lot of questions or test about anything. It starts to begin the big centres nowadays but now I don't know that where we should go and ask about anything, about genetics.'*

TT040: *'Oh actually I do know that there are some organisations specialised in that,' [test taker elaborates on the topic here], I really don't know about the details, but there is one thing that I do know that actually Iranian people are not really that concerned about genetic research'.*

Q: How do people in your country feel about genetic research?

TT009: *'(U:m) I (.) I really don't, don't hear about the people about genetics here around, but my sister in America, she said a lot of things about genetics, but in Iran we couldn't say a lot of things about genetics'*

TT017: *'Mm (.) this is, this is not a common, uh, this is not a common subject in media or uh in government uh policy, so I guess no one, most of the people do not aware of genetic research.'*

TT032: *'Mm I don't think they are really interested in that. Not because I'm not, because I haven't seen people around me that are interested in that.'*

TT043: *'Um mm I don't believe that they care about it so much, um I don't think they are um, maybe that's my opinion but from the people around me I can say that none of them is into these things, researchers about genetic'*

TT069: *'Um in my country genetic research is not so popular, and is not so mm famous among people, but if they want I think they have to um ask it from like clinics and hospitals or something like this. (69)*

TT074: *'I think uh people uh most people in my country don't pay attention about uh this uh this topic' (74)*

The extract below suggests that the unpopularity of the topic might be tied to religious reasons:

TT027: *'Uh I think some of uh some of the people in my country, because of uh something in our religion, they don't like lots of things in genetic research. For example all of them uh they think uh we are created from [soil]. And they can't uh they don't like some researchers that says it's not true, because of that I think they don't like it.'*

The mismatch between topic and the social, cultural and religious backgrounds of the test takers in the study is most perceptibly observed in the Festivals and Genetic Research topics. However, similar comments, which allude to the distinctive aspects of the Iranian society and thus render certain topics or questions fairly irrelevant, were also observed in relation to the topic of 'Dancing' (A.6):

Q: Tell me about any traditional dancing in your country.

TT046: *'In my country actually we had some kind of traditional dancing that they danced, you know, with each other, for example in a group, I mean they doing something like each other, and with beautiful music, beautiful I mean costume, and uh you know beautiful I mean movement, but these days I cannot find, in my country, these kind of dancing, because it is a little bit forbidden.'*

TT058: *'Uh well because of the religion, the religious boundaries that we have we don't usually let these things be improved in the country, be like, you know, grow in the country, but we see in small places that like in weddings that they come and start the traditional dances, and just uh enjoy themselves, but it's not so you know common to talk about it.'*

Q: Do you think that traditional dancing will be popular in the future? [Why/Why not?]

TT017: *'Uh it really depend on the government, because in Iran dancing is forbidden, so if the country uh let the people dance maybe the traditional dancing be popular, if they don't it will be forgotten totally'.*

Q: Has anyone ever taught you to dance? [Why/Why not?]

TT002: *'Um yes, once I have, I have um take some salsa class, and but because we live in a country that dancing is not allowed, especially for the women, so that you have to learn it by yourself. And maybe you have to learn it from the TV shows and something like that'.*

These extracts demonstrate that relatively speaking, some topics may not be appropriate for particular countries and cultures. In other words they may lack 'local' validity (O'Sullivan, 2011). However, this is not to say that these topics were problematic for *all* test takers. Some test takers managed to elaborate on these topics without exhibiting any content-related problems. The following extracts illustrate how information-based content can allow for topic development:

Q: Where do people in your country find information about genetic research?

TT058: *'Um one of the things of course is internet, using the internet, there are some uh special courses in Tehran University and other places, some special organisations, that have now uh lectures uh about uh genetic human geno project, and also it's not so relative but [00.27] technology related to the branch of genetic in fact. And recently I've heard that in some pre-school years even it's working, they are working about that genetic things to children be you know just um familiar with the topic'.*

Q: How do people in your country feel about genetic research?

TT072: *'I think people use this topic as just a conversation opener or just as a topic to talk about during lunchtime, uh I don't think people care very much about it, especially I know it's a hot topic, especially about the ethics, uh whether it's right or not, in some other countries? But I don't think Iranians care about it yet very much'.*

TT023: *'Um these days I think it's getting more popular in the country, especially if you go to some uh institute for example sonography, which I did last year for my baby, they have one room for the genetic, you go and you talk about your family, if you have any diseases in your family, if you have any relative married in your family then they draw on tree and they tell you if you have problems or not, or if something critical or something might happen. So they give you that um that um calmness that nothing is going to happen, or alarm that you have to be aware of this, so we have to take uh more tests to see whether the baby inside is OK or something is wrong'.*

The availability of background knowledge to the test takers in the above extracts serve to illustrate the inadequacy of making assumptions about topic familiarity for all test takers on the basis of a group-level factor such as cultural background. The level of background knowledge of a person might reflect an interaction between a topic and a whole range of other factors. For example, one of the raters highlighted the role of age and in performances on topics, noting that some of the topics might be to be favoring more mature test takers, as they would arguably have more experiences and interests to draw from. It is therefore important to establish the degree of background knowledge empirically instead of relying on stereotypical notions of familiarity, which may or may not be applicable to all individuals within a group.

6.2.8 Topic Interactions with Characteristics of the Individual Test Taker

The previous section showed that some topics and/or questions might prove problematic for the majority of individuals within a sub group of the test taking population due to the mismatch of the topic with the social and cultural background of the test takers. The purpose of this section is to show that such a mismatch can also be exhibited at a more individual level where a) lack of personal interest in a topic, or b) a negative affective influence of a topic might inhibit performance to a certain extent. Examples of each category are provided below, although it should be noted that there are, at times, a degree of overlap between the factors as well:

Lack of personal interest or experience with the topic

Q: Tell me about any traditional dancing in your country.

TT008: *'Uh I have to be honest with you I am really not into traditional dancing and traditional music'*

Q: What do you enjoy most about it [Festivals]?

TT012: *'Actually I didn't ever enjoy the festival [01.00] uh people think that they can enjoy the [01.11] and they can show things to people around the [01.14] so people, maybe some people do enjoy that thing, but I don't actually enjoy it.'*

TT030: *'Um nothing special for me in this festival, it's not attracts me.'*

Q: Describe one of your friends.

TT003: *'OK this is difficult topic for me because I haven't a special friend but what I can say for imagine my friend' 'I cannot explain for you because I haven't a special friend I cannot explain but this is my idea.'*

Negative affective influence of topics

Q: How does water transport, like boats and ships, compare with other kinds of transport?

TT029: *'Well actually I'm not very good at that because I don't feel well uh in the boats, and (.) I think it's a bit um danger, more dangerous than other transports, uh that's why I am a little bit afraid of the sea and the boats as well.'*

Q: Do you enjoy dancing? [Why/ Why not?]

TT017: *'Because I'm not a good dancer, and I feel a little bit stupid when I dance.'*

TT038: *'I will feel ashamed and I'm not so comfortable when I want to dance.'*

Q: Has anyone ever taught you to dance? [Why/Why not?]

TT001: *'No, because I never wanted to learn it. It's not that I don't like it, it's like I like it but I know I can't do it. [00.17] like specific things in life that you know you are sure you can't do. One of those things for me was dancing, I knew from the time I was ten years old I couldn't be a good dancer, so I never went for it.'*

TT017: *'Uh I uh take a few courses in dancing in Arabic dancing, but the atmosphere was not very good, and I tried to show my (00.43 EITHER – heart – OR – art) in dancing to my husband but he humiliated me (...) I didn't attend the course any more.'*

TT038: *'I will feel ashamed and I'm not so comfortable when I want to dance.'*

In these examples it is the topic of 'Dancing' which is most often associated with negative affective schemata. The following excerpts, in contrast, are suggestive of the positive affective influence of the same topic

TT046: *'Uh because as I lose my energy I get more energy back, I mean it's full of fun, it's it's not like, I don't know, studying, because I hate studying, but it's like something that you have your own experience, and you can make your own choice, decision I mean, that for example going left or going right, or doing that, or doing this, and this is really makes you happy.'*

TT002: *'Because it gives me some power that I think I can um (.) um explore the world. [LAUGHS] I don't know why, I really love to dance.'*

Taken together these examples illustrate the distinct ways in which topics interact with individual characteristics of the test takers. An examination of the rater comments also suggests that raters may also be interacting with the content of what the test takers are expressing. One rater compared the speaking tests to a small window into the culture and personal lives of the test takers while another rater compared the rating process to 'speed dating' where one gets to know something about different people in a very short amount of time.

6.2.9 Test Constraints and Lack of a Support Mechanism

The final theme in this chapter touches on general aspects of the test which might magnify the perceived impact of topic-related problems on spoken performance. The raters listed the standardised nature of the test, the use of an examiner script and the asymmetrical power relations between the test taker and the examiner as problematic features of the test which can impede both examiners and test takers from attempting to repair problematic topic sequences. To illustrate, the examiner script may not allow examiners to easily shift topics (in case of a problem) or to provide support to the test takers. The test takers, on the other hand, are also not in a position to shift topics as the interview is entirely managed and driven by the examiner. The artificial nature of the test and the lack of a support mechanism for test takers are captured in the examiner comments reproduced below.

Rater Extract 18

The nature of interaction is very contrived. It's really really unnatural and weird and sometimes painful to listen to because you just feel it's just you feel like that the candidates really struggle with the situation it's just so unnatural that um because they get very little verbal feedback and there isn't any back channeling coming from the interviewer and so just the natural I mean it is actually inhumane. It's true that it is standardised and I can understand the thinking behind it and I hate to say this (.) it's actually (.) um if that's gonna be the case it's better to give a speech sample à la TOEFL or PTAE because you might as well be a machine because you know we're supposed to communicate you're breaking a fundamental part of what makes up human, human communication, actual interaction.

Rater Extract 19

I mean examiners have strict rules about what they can and can't do. I mean my instinct would be to give some examples but the questions themselves don't necessarily... I mean don't elicit speech and people's imagination (.) this is the thing that I as a teacher I find difficult, um that you have some students who are very articulate but they don't have great imagination or once you get them talking they can talk very fluently but they can't think of the ideas so unless you have some sort of a brainstorming activity they can't really think of anything.

In extract 19, the rater draws from her teaching background to highlight the importance of prompts or brainstorming activities for eliciting speech. Interestingly, the results of this study also suggest that Task Type B topics, in which test takers are presented with such prompts, exhibit the least content-related problems: only three performances out of the 121 selected for qualitative analyses were Task B topics. It can therefore be argued that the Task B topics, by providing information points for test takers to talk about, can level the playing field in terms of background knowledge and thus minimise the negative effect of topic unfamiliarity.

The discussion on this theme has highlighted some of the negative features of the test which can hinder test takers and raters from repairing problematic topic sequences which might be due to lack of background knowledge while also singling out those test features which can reduce the background knowledge effect. These findings can have important practical implications and are revisited in the next chapter.

6.3 Summary

The qualitative analyses of rater interview data and the content of test taker response data revealed a number of interesting findings into the ways in which test takers deal with topic unfamiliarity from abandoning the topic to adopting a variety of test taking strategies. Lack of background knowledge was also discussed in relation to potential impacts on test takers, their performance and scores as well as raters' decision-making process in rating performances. The insight from the analyses provided possible explanations for some of the quantitative findings of the study, specifically in respect to the absence of a practical effect of topic and background knowledge of topics on scores. The findings also focused on problematic aspects of certain topics which rendered them irrelevant to a group of individuals as a function of social, cultural or religious backgrounds. Attention was also drawn to the complex ways in which individual characteristics of test takers can interact, in distinct ways, with different topics thus illustrating the inadequacy of making assumptions about test takers' levels of background knowledge. Lastly, the issue of a support mechanism for test takers in dealing with problematic topics was discussed in relation to different test features.

Chapter 7 Discussion

The aim of the present study was to examine, from a test validity perspective, the effects of topic and background knowledge of topic on speaking performance in language proficiency interviews. The study's main overarching research question was formulated as follows:

How is the topic validity of a test of second language speaking performance influenced by the random assignment of topics to test takers who bring different levels of background knowledge to the topics?

The study's main research question delineates the two variables of topic and background knowledge of topic as factors which can have an impact on performance and can potentially introduce construct-irrelevant variance into speaking performance assessment. The systematic variation in topics of a speaking test, exercised through the random assignment of topics to test takers, reflects an underlying assumption of topic equivalence in terms of difficulty and the comparability of resulting test-taker performances. It also implies that differences in the levels of background knowledge that test takers bring to these topics as a function of group-level factors (e.g. cultural background) and/or individual test-taker characteristics (e.g. interest in topic, experience with the topic) do not exert a systematic and large influence on test-taker performances. Evidence to the contrary would signal the presence of construct-irrelevant variance and test bias. It was also posited that background knowledge might differentially interact with test takers' proficiency levels. Moreover, the rater-mediated nature of performance assessment indicates that in

examining the effects of the above variables on performance scores, it is necessary to control for systematic differences in raters' scoring behaviour on the one hand, and to inquire into potential interactions between characteristics of the raters and the variables under investigation on the other.

In addressing the foci of the study, a mixed-methods strategy of inquiry was adopted (Creswell, 2003; Creswell and Plano-Clark, 2007) which allowed for an examination of the research problem through different lenses. The quantitative data (from speaking test scores, background knowledge questionnaires, C-test scores, functions observation checklist, test-taker close-ended questionnaire responses) were explored using MFRM analyses, multiple regression and descriptive statistics (discussed in detail in Chapter 5). The raters' interview transcripts and a sample of the test-takers' content of speech were analysed qualitatively (see section 6.1 for details of the qualitative steps taken). In this chapter, the findings from the quantitative and qualitative research strands are summarised and synthesised thematically and in relation to the study's overarching research question. Convergent and divergent findings are delineated and discussed in more depth. The study is also positioned in the wider literature by comparing and contrasting its findings with other empirical research and theoretical frameworks. The findings are subsequently integrated within Weir's (2005) socio-cognitive framework of language test validation in building a topic validity argument. Next, the implications of the study's findings are presented. The chapter ends with a consideration of the study's limitations and directions for future research.

7.1 Summary and Synthesis of Findings

The findings from the study's quantitative and qualitative strands of inquiry are brought together and summarised in this section. It should be noted that the study's eleven subsidiary research questions are not independently presented but are discussed in relation to five broader themes as follows.

7.1.1 Topic Effects on Performance

The eighteen topics used in the study were quantitatively compared both at the test (form) level and at the task (type) level. Results showed that at the test level, the topics in the study exhibited difficulty measures which were statistically distinct (see section 5.3.5). This was as expected, in light of the fact that the tests are designed to include task types of differing difficulty levels.

When focusing at the task (type) level, results indicated that topics within each task type can be generally divided into a minimum of two statistically distinct difficulty levels and therefore cannot be considered equivalent or 'parallel' (see section 5.3.8). These differences in topic difficulty measures were then examined in relation to the average speaking ability necessary to move across adjacent band scores and across criteria. Findings illustrated that these differences were not large enough to translate into differences in test-taker scores (see sections 5.3.8.4 and 5.3.8.5). In other words, the minimum speaking ability required to move across band scores consistently exceeded the maximum difference between topic difficulties, both at task and test levels. The results demonstrate the absence of a practical effect of differences in topic difficulties on performance scores in this study and thus serve as an element of topic validity evidence

which justifies, on the basis of the score data, the random assignment of topics to test takers in the context of the test investigated here.

The equivalence of topics (at the task level) was examined more qualitatively by focusing on the comparability of the range of language functions elicited across different topics using an observation checklist. To remind the reader, the functions observation checklist (O'Sullivan, Weir and Saville, 2002: 38) was originally designed as 'an effective and efficient procedure for establishing the content validity of speaking tests'. The checklist was adapted for use in this study to examine the extent to which systematic variations in topics of the tasks (within each task type) influence test-takers' performances in respect to the language functions listed by the IELTS test developers as representing the construct of spoken language ability (O'Sullivan, Weir and Saville, 2002). In this study, illustrative comparisons of the observed functions in two topics within each task type revealed both similarities and differences in the range of functions that topics elicited (see section 5.3.9). These variations in observed functions can be interpreted to imply that different topics can potentially tap into different aspects of the underlying speaking construct and as such, cannot be considered comparable, regardless of score stability. In other words the inferences drawn on the basis of the same score on two different topics may not be valid, as scores may have different meanings corresponding to the specific language functions they represent (O'Sullivan, Weir and Saville, 2002). Two important caveats need to be taken into account in relation to this interpretation. Firstly, in the operational IELTS Speaking Test a given score represents performance (and language functions) across five topics and it is therefore likely that variations in observed functions balance out across the whole test. Secondly, as Weir and Wu (2006: 177) point out, the discrepancies in observed

functions across topics might reflect differences in the ‘proficiency’ of the test takers responding to the topics and not necessarily ‘the variation in the coverage of language functions between (...) task versions’. These issues were not examined in the current study and require further inquiry. The study’s findings are therefore not considered conclusive but rather suggestive of topic effects on qualitative differences in the language produced by test takers.

Taken together, these findings strongly resonate with Fulcher’s (2003) position in relation to the potential effects of task conditions on performance. Recall that in Chapter 2, the SLA standpoint (Tarone, 1988;1998), which held that task conditions can have an impact on discourse and consequently on performance scores, was challenged by Fulcher (2003) and Fulcher and Marquez-Reiter (2003) who questioned the assumption that changes in candidate discourse, as a result of differences in task conditions, ‘*automatically translate into changes in test score*’ (Fulcher, 2003:64 *italics in original*). The current findings are clearly in line with Fulcher’s (2003) argument: while the results of the analyses of functions suggest qualitative differences in the language produced by test takers attributable, in part, to the task condition of topic, the MFRM analyses of score data illustrate how statistically significant differences in topic difficulty measures have failed to have a practical effect on test scores.

7.1.2 Background Knowledge Effects on Performance

The quantitative results of the study generally suggested a statistically significant effect for BK with a small effect size, failing to reach practical significance (see sections 5.4.2, 5.5.4 and 5.5.5). The impact of topic unfamiliarity on the language produced by test takers was also one of the emergent themes from the analyses of rater interview transcripts

(section 6.2.4). This brings the discussion to an evaluation of the impact of background knowledge on spoken performance. Both parallels and contrasts were observed across the quantitative and qualitative analyses of data. The results converged in respect to background knowledge exerting an impact on performance but diverged in relation to the *degree* of this impact. These results are explored in more detail below.

The raters in the study noted an adverse impact of topic unfamiliarity on test-takers' performances. Similar findings are drawn from the results of the MFRM analyses where BK was parameterised as an additional facet of assessment and was shown to have a statistically significant main effect on performance. This is compatible with the rater observations, as a consideration of Fair-M average results across the different background knowledge conditions (Low, Medium and High BK) showed that when relative speaking proficiency of test takers and relative severity of raters have been accounted for, test takers exhibit lower average performance measures in the Low BK condition compared to the Medium and High BK conditions. These differences in performances across BK levels were shown to be statistically significant and low levels of background knowledge (suggesting unfamiliarity) were shown to consistently and systematically pose the greatest level of challenge for test takers. By the same token, high levels of BK are suggested to have a facilitative effect on performance. These findings are in line with the results of studies which have found a statistically significant role for background knowledge in second language performance (e.g. He and Shi, 2012; Huang, 2010; Krekeler, 2006; Schmidt-Rinehart, 1994; Tedick, 1990) . In contrast, the current study's findings run counter to the findings reported in Jennings et al. (1999) and Lee and Anderson (2007) where background knowledge, operationalised as participants' 'choice' of topics and

‘departmental affiliation’ respectively were shown to have non-significant effects on writing performance scores. The integrated nature of the assessment context in both these studies can serve as a possible explanation for the absence of a significant background knowledge effect. In the words of Jennings et al. (1999:448), ‘the context provided by the test materials, had reduced the impact of prior knowledge to the point of insignificance’. The statistically significant impact of BK on performance in the current study can thus be a reflection of the independent nature of the speaking tasks (particularly task types A and C) which require test takers to rely on their own background knowledge in responding to questions.

It is on the extent to which statistical significance translates into practical significance (in terms of influence on achieving higher or lower band levels) that the results from the different analyses diverge. The analyses of test takers’ questionnaire data indicate that the majority of test takers (95%) place a great importance on both topic and background knowledge variables as factors which can have a major effect on their final scores. This finding echoes the results of other empirical research which have looked at test-taker perceptions of topic and BK effects on scores (e.g. Jennings et al., 1999; He and Shi, 2012). In a similar vein, the raters in the study attributed an important role to background knowledge, or lack thereof, in affecting features of test-takers’ performances on the one hand, and scores awarded to test takers on the other (see section 6.2.4). In relating performances to the rating scale criteria, the influence of topic unfamiliarity was often associated with an increase in pauses and hesitations (FC scale), a decrease in lexical diversity (LR scale) and grammatical complexity (GA scale) and inferior topic development (TD scale). A drop in the range of observed functions was also commented

on. Pronunciation (P scale) was pointed out as the scale least likely to be negatively affected by lack of background knowledge. In contrast, an adverse effect on ratings for the remaining scales was predicted.

The quantitative analyses of score data (from MFRM and multiple regression) ran counter to the results of test-taker perceptions and rater interview transcripts, as the statistically significant difference between different BK measures failed to translate into practical (meaningful) effect on test scores. MFRM findings showed that the maximum difference between the least challenging (High BK) and most challenging (Low BK) BK levels fell below the minimum average speaking ability required to move across adjacent bands for the different criteria. The absence of a large (practical) effect of BK on scores was further substantiated in the multiple regression analysis of data where BK was found to be a significant predictor of speaking proficiency scores, but only explaining 3% of the variance. As expected, general language proficiency exhibited a much stronger predictive power, accounting for 60.1% of the variance. One can argue that the quantitative findings from the score data should perhaps be given more weight in this discussion, as people may not be able to reliably evaluate the actual magnitude of BK.

According to the MFRM results, the only criterion for which BK was found to have the potential to exert a large (practical) influence on scores was the TD criterion (but only at specific categories on the scale). The results of the multiple regression analysis also showed that once TD was removed from the analysis, the predictive power of BK was reduced from 3% to approximately 1% (albeit still significant). The sensitivity of the TD criterion to differences in BK levels is partially supported in Lumley and O'Sullivan's (2005) study of the effects of gender-oriented topics on speaking in which the authors

report a small advantage for males on the ‘task fulfillment and relevance criterion’ for those topics which males were assumed to have higher background knowledge of (*Ibid.*: 432-433). These findings also imply that the TD criterion is functioning as intended in absorbing BK effects.

Broadly speaking, the quantitative findings have suggested a statistically significant effect for BK with a small effect size, failing to reach practical significance. BK was shown to account for only 1-3% of the variance in speaking scores while general language proficiency exhibited a much stronger predictive power, uniquely accounting for approximately 60% of the variance. These results are in line with empirical findings from various other studies. For example, in Jensen and Hansen’s (1995) study and in the context of listening assessment, prior knowledge was found to have a statistically significant main effect yet small effect size on listening performance, accounting for 3-9% of variance in scores on a subset of lectures. The small proportion of variance explained by topic-related factors compared to the variance explained by general language proficiency was also reported in Papajohn’s (1999) study on speaking performance where 4.7% and 67.2% of score variance was predicted by topic groupings and general language proficiency respectively. Within a task-based performance context, Skehan et al. (2012) reported significant effects for topic familiarity in terms of fluency, accuracy and lexical sophistication measures (in line with the rater perceptions in the current study), although the effect sizes were found to be small. The negligible effect sizes led the authors to conclude that ‘speaking about something one is familiar with does produce performance advantages in (...) various measures, but the advantage is surprisingly small’ (Skehan et al., 2012: 178). Reflecting these findings, the general trend that background knowledge

has a significant yet negligible influence on performance scores is supported in the present data.

What remains to be addressed is an explanation for the divergent results touched upon earlier in terms of the mismatch between the perceived (strong) influence of topic and BK on speaking scores (as voiced by test takers and raters) and the absence of a practical influence of these variables on scores (as observed in the quantitative strand of findings). Possible justifications for this disparity are presented below.

One of the qualitative themes of the study pertained to the constraints of the test and the absence of a built-in support mechanism for dealing with topic-related problems (see section 6.2.9). Another theme touched on the format of the speaking tasks as potentially mediating the effects of background knowledge of topic (discussed in section 6.2.6). Building on these themes, it is argued that certain features of the test and test tasks can simultaneously magnify the salience of topic-related effects while also moderating the (negative) impact of lack of background knowledge.

In Chapter One, two important features of the IELTS Speaking Test in relation to topics were outlined: firstly, the centrality of topic in generating speech and secondly, the use of topic as an organisational tool for managing the interview, constrained and standardised by the examiner script. On the basis of the qualitative themes of the study, it is now postulated that the combination of these two features magnify the perceived impact on topic/background knowledge of topic influence in the following ways. On the one hand, the dominance of topic as the main elicitation tool can render topic as the most salient feature of the speaking test. The independent nature of the speaking tasks also implies that test takers have to rely on their own background knowledge in responding to the different

topics. On the other hand, the standardised nature of the test, the governing role of the examiner script in guiding the interview, and the power imbalance between examiners and test takers indicate that in cases where a topic-related problem arises in the test, neither test takers nor examiners feel that there is room for addressing the problem (Seedhouse and Harris, 2011); for examiners, because they are constrained by the test format and script, and for test takers, because they feel they have little to no control over the choice of the topic or the direction towards which the interview is going. In the words of one of the study's raters, the rather 'contrived' nature of the test may leave the parties with little to fall back on in case of problems to do with a mismatch between the test developers' intention of generating samples of speech through topics and an undesired outcome where the fulfillment of the original intention is obviated by a construct-irrelevant factor such as the candidate's lack of background knowledge of the topic. This is one possible reason for the perceived significance of topics and topic unfamiliarity by test takers and raters alike. A second reason alluded to by the raters in the study pertains to the potential emotional impact of topic unfamiliarity on test takers in inducing negative reactions such as anxiety, confusion or apprehension which can arguably leave a strong affective impression on test takers (Bachman and Palmer, 1996) and thus explain the prominence of topics and background knowledge of topics from the perspectives of the test takers.

Three reasons are now proposed as a rationalisation for the absence of a strong background knowledge effect, despite its perceived salience in influencing scores: 1) the inclusion of the TD criterion, 2) the multi-question, multi-task format of the speaking tests and 3) test-taking strategies. These reasons are expanded upon below.

The first reason relates to the inclusion of the TD scale. It is proposed here that TD served as the obvious choice for capturing the more non-linguistic features of test-takers' performance in respect to the development of ideas and for absorbing the background knowledge-related effects, thus reducing the influence of BK on the remaining criteria. Support for this argument is drawn from the quantitative results of the study which isolated TD as the only scale likely to be strongly affected by differences in BK levels.

The second reason pertains to the multi-question format of the speaking tasks. Insight from rater remarks and the analyses of the content of test-taker performances demonstrated that it is not necessarily a topic which might be unfamiliar or problematic but rather, specific questions within the sequence of topic-based questions which might require test takers to draw on background knowledge that is not available to them; an observation which is in line with findings by Seedhouse and Harris (2011) in relation to problematic questions in the IELTS speaking tasks. To further elaborate, BK might exert a strong influence at the question level within a topic sequence but the multi-question format of the task serves as a control mechanism for reducing the potential impact of lack of BK, as speech is likely to be generated by other follow-up questions on the same topic thus allowing raters to apply the rating scales reliably to the performance. By extension, the multi-topic format of the speaking test further safeguards against the negative impact of BK on scores, as the availability of a minimum of five topics at the test level ultimately reduces the likelihood of *all* topics being unfamiliar. Any (negative) influence of BK on linguistic features might therefore fade within the broader performance. One can raise the following question in this respect: should the potentially negative effect of a lack of background knowledge on the test taker, examiner (and/or rater) and the quality of

performance, albeit limited, be ignored in light of the finding that background knowledge, in effect, does not have an impact on final scores?

Lastly, attention is drawn to the role of test-taking strategies as another mediating factor which can potentially reduce the impact of lack of BK on performance scores. As evidenced in rater remarks on the one hand, and in illustrative examples from the content of test-taker performances on the other, some test takers can circumvent BK-related obstacles by deploying different test-taking strategies (e.g. speculating, waffling or going-off on a tangent) which allow them to generate speech regardless of topic unfamiliarity. Interestingly, raters' perspectives on the use of such strategies were mixed: on the one hand, the speech-generating function of these strategies was positively viewed, as it allowed for rating of linguistic features of performance. On the other hand, the use of strategies such as waffling or going-off on a tangent was negatively perceived due to test takers' failure to address the task adequately, leading one rater to question whether there is any real 'coherence' in a response that does not answer a question, no matter how extended that response is. On the basis of these findings, it can be inferred that the lack of topic-related BK may instigate different test-taking strategies which have the potential to minimise the impact of lack of BK on the amount of generated speech and linguistic features of performance though not necessarily on the fulfillment of the task in terms of topic development.

7.1.3 Background Knowledge and General Language Proficiency

The quantitative and qualitative results of study generally converged in relation to the interaction between background knowledge and general language proficiency and their respective contributions to overall performance.

The results of rater interviews suggested that lack of background knowledge might differentially affect test takers from high and low proficiency levels. This observation was independently and empirically addressed in the quantitative analyses of the study. The results of a bias analysis between background knowledge and proficiency indicated that when background knowledge levels are low, examinees with low proficiency levels are at a disadvantage compared to high proficiency examinees. An opposite effect was not observed. In other words, high degrees of BK do not advantage/disadvantage examinees from different proficiency levels. It is only when BK is low that high proficiency test takers have an advantage over test takers from low and medium proficiency levels. Insight from rater remarks suggests that low proficiency examinees are more likely to be negatively influenced emotionally (e.g. frightened or confused) by unfamiliar topics, leading them to abandon the topic or disengage from the question whereas higher proficiency test takers may be able to adopt a variety of strategies (as explained in the previous section) to deal with problematic topics.

The results of the test-taker questionnaire responses also partially align with the above findings as evidenced in the shift in the pattern of responses of the participants when they were asked to express their views on influence of topic and BK of topic on scores. The majority of the participants (95%) agreed that choice of topic and having ideas about a topic might affect their final scores. However, when the same statement was revised to include a proficiency element, that is, ‘choice of topic is not important if my English is good enough’ only 62% of the participants disagreed with the negatively worded statement with a 33% agreement rate. It can therefore be inferred that, at least in the

perspectives of some test takers, topic and BK of topic are likely to exert less of an influence when language proficiency is higher.

The relative contribution of background knowledge and general language proficiency in explaining variance in spoken performance was covered earlier. To reiterate, BK was shown to account for 1-3% of the variance in performance scores on topics which, in comparison to the 60% of variance explained by general language proficiency, is very small and negligible. The MFRM results showed that the distribution of examinees speaking ability levels is much wider than the distribution of different BK conditions (and topic difficulties). It can therefore be inferred that higher order constructs such as general language proficiency and/or oral speaking proficiency are the main determinants of performance in the test under examination (note that a strong positive correlation was observed between C-test results and speaking ability estimates). These findings run counter to the conclusions drawn from He and Shi (2012:460) in respect to the influence of prior knowledge on impromptu essay writing performance in which they '[reject] the assumption that language proficiency is the main factor determining performance'. In contrast, the quantitative findings in the current study are in line with the qualitative views expressed by the test takers in Huang's (2010:221) study on the influence of topical knowledge and anxiety on spoken performance in integrated and independent speaking tasks where 'participants asserted that overall oral proficiency (...) outweighed topical knowledge and anxiety in terms of impacting oral test performance'.

Additionally, the findings from the results of two rater bias analyses (rater x background knowledge and rater x general language proficiency) showed that the raters in the study did not differentially rate the performances of test takers from higher or lower levels of

background knowledge. However, they exercised distinct types of bias towards test takers from different proficiency levels (either exhibiting more leniency towards low performers but more severity towards higher performers or vice versa). While these results are only based on four raters, they nevertheless highlight that it is general language proficiency of test takers – and not relative background knowledge – that interacts more strongly with the characteristics of the raters.

Taken together, these findings suggest that general language proficiency and oral speaking ability are the main determinants of performance on the speaking tests (reflecting the purposes of the test) and that the influence of the construct-irrelevant variable of background knowledge of topic on speaking performance scores is minimal, despite its statistical significance. These can serve as important pieces of topic validity evidence for the speaking test under examination.

7.1.4 Background Knowledge and Topic Difficulty

The study's fourth important set of findings relates to the interaction between topic difficulty and levels of BK where, once again, the quantitative and qualitative strands of inquiry intersected.

The results of the MFRM analyses in examining the influence of BK on the measurement results of other facets revealed that, with the exception of the topic facet, BK had virtually no impact on the other facets' measures. A comparison of topic measurement results (when BK was included/excluded as an additional parameter in MFRM) revealed that the influence of BK was manifested in a) shifts in the ranking order of different topic elements, b) changes to topic difficulty measures and c) variations in the overall range of topic difficulty. In other words, unlike the other facets, the measurement results of the

topic facet fluctuated with the inclusion of the test-taker characteristic of background knowledge of topic. These fluctuations were not substantial. Nevertheless, the finding that the measurement results of the topic facet did not remain stable is a very important one; it lends strong support to the argument advanced by Bachman (2002:464) as discussed in Chapter 2 and repeated here, in which he problematises the conceptualisation of the notion of ‘difficulty’ as residing exclusively in the task:

I argue that most ‘difficulty features’ (...) are not inherent in tasks themselves, but are functions of the interactions between a given test taker and a given test task. Next, I argue that empirical estimates of task difficulty are not estimates of a separate entity, ‘difficulty’, but are themselves artifacts of the interaction between the test-taker’s ability and the characteristics of the task.

The observed variations in topic difficulty measures in the analyses with and without BK illustrated the sensitivity of topic estimates to the amount of background knowledge that test takers bring to the topic. This was particularly striking for two of the topics (Festivals and Genetic Research) which, when BK was not parameterized, were estimated as the most difficult topics within their respective task types (A and C). However, when BK was modeled in MFRM, the difficulty estimates of the topics changed, exhibiting lower difficulty measures. In other words, it was not the topics *per se* that were difficult but rather, the lack of background knowledge in this particular sample of test takers that interacted with the topic, resulting in higher difficulty indices.

The qualitative analyses of the content of test taker performances substantiated and shed further light on the quantitative results by identifying the same two topics of Festivals and Genetic Research as culturally unfamiliar for the Iranian test takers in the study. Questions on these topics required test takers to draw on previous experience or knowledge which they, as a group, did not necessarily have available on account of their Iranian background and other cultural and religious factors. Illustrative extracts from test performances

displayed how the test takers found these specific topics irrelevant to their local context. It can therefore be argued that these topics lacked ‘local’ validity (O’Sullivan, 2011).

Moreover, the qualitative results displayed several instances of question-answer sequences where the unfamiliarity of the question or topic resulted in candidates closing the question by providing a short response such as ‘I don’t know’ or ‘I’m not really interested in that’ or ‘we don’t have any festivals in our country’. One of the raters in the study reported facing a dilemma in rating such performances, as she found the responses to be ‘perfectly adequate’ in light of the test takers’ lack of background knowledge but nevertheless unacceptable in terms of topic development. This finding is in line with the observation made in Seedhouse and Harris (2011) in respect to the priority of the question-answer component (of a speaking task) to the topic component in occasions when the two components did not coincide, leading the authors to contend that ‘candidates can answer questions without developing topics’ (*Ibid.*: 73). On the basis of this study’s findings, it can be argued that lack of BK or topic unfamiliarity may increase the likelihood of candidates responding to questions without developing or elaborating on topics, which is arguably against the speech-generating function of these questions and topics.

Another parallel can be drawn between the qualitative findings of the current study and those by Seedhouse and Harris (2011) in relation to problematic questions or topics within the context of IELTS speaking test:

Problematic questions may involve an unmotivated shift in perspective, may require specialist knowledge of experience which may not be available to most candidates, or may be puzzling in some way (...) A sequence of questions on a particular topic may appear unproblematic in advance of implementation. However, this may nonetheless be a cause of unforeseen trouble for candidates.

Seedhouse and Harris (2011:105)

The results of the current study clearly showed that a seemingly familiar topic such as ‘Festivals’ was a source of ‘unforeseen trouble’ (Seedhouse and Harris, 2011:105) for most of the Iranian test takers in this study. On the other hand, contrasting extracts were used to demonstrate that even within the same local context, some test takers managed to elaborate on the same problematic topics without exhibiting any content-related problems. The combination of these findings serves to illustrate the way in which a given topic can potentially bias against a particular group of test takers but also demonstrate the inadequacy of making stereotypical assumptions about topic familiarity for *all* the test takers within that group on the basis of a group-level factor such as cultural background.

The qualitative results also exemplified the intricate ways in which a test taker engages with a topic. The task topics were shown to interact with characteristics of the test taker such as their previous experiences, gender, affective schemata, cultural and religious backgrounds and personal interests, illustrating the need to establish the familiarity or difficulty of a topic by going to the level of the individual test taker. These findings echo Lumley and O’Sullivan (2005:432-433): ‘task difficulty is too complex to be categorized in (...) simplistic terms (...). Tasks are more likely to affect individuals differentially’.

7.1.5 Role of the Topic Development Criterion

To remind the reader, the topic development criterion was included in this study as means of isolating the content-related effects of topic and background knowledge of topics on performance scores. All quantitative analyses were therefore carried out both with and without the TD criterion. In this section, the findings from these analyses are brought together in order to evaluate the contribution of the criterion within the assessment context.

The results of the quantitative analyses showed that the TD scale was the easiest of all five criteria, exhibiting a difficulty level which was markedly lower than the other criteria. Test takers have therefore an increased likelihood of achieving a high score on the TD criterion compared to the other criteria. The pronounced sensitivity of the criterion to differences in topics and background knowledge of topics implied that the scale was functioning as intended in absorbing the more non-linguistic and content-oriented features of the performance. Its inclusion appeared to inflate the topic difficulty range while decreasing the examinee speaking ability range.

The separation reliability values indicated that the criterion reliably separated test takers into statistically distinct ability strata. The criterion's fit statistics, while relatively high, were nevertheless within an acceptable range, indicating that the criterion was contributing, as an independent criterion, to the underlying speaking construct and thus explaining some of the variance in the speaking scores. On the other hand, an examination of the table of unexpected responses showed that the criterion was associated with large residuals, mostly as a function of low ability examinees achieving scores higher than expected on the TD criterion and conversely, high ability examinees achieving scores lower than expected. These findings generally suggest that, despite some limitations, the TD criterion can contribute uniquely to the assessment context and its inclusion needs to be evaluated in light of test purposes.

7.1.6 Psychometric Quality of the Speaking Test

The final set of findings pertains to the psychometric properties of the speaking test under examination. The MFRM results showed that the speaking tasks in the study exhibited a range of difficulty levels and were successful in distinguishing between test takers from

different abilities. The criterion facet results suggested that the five criteria in the study exhibited statistically distinct levels of difficulty and contributed in distinct ways to the separation of test takers into different ability levels. An examination of the rating scale structures and categories demonstrated that the categories within the scale were generally functioning as intended. The effects of the other facets of the speaking assessment context (raters, topics, background knowledge of topics) on scores were shown to be very small in comparison to the speaking ability levels required to receive higher or lower band scores across the different criteria. The overall fit of the speaking score data to the Rasch model was also satisfactory. Taken together, these quantitative findings suggest that scores on the speaking tests predominantly reflect the underlying speaking ability construct that the test was designed to measure and provide strong evidence for the (topic) validity of the test under examination.

7.2 Towards Building a Topic Validity Argument

The study employed Weir's (2005) socio-cognitive framework of language test validation (discussed in detail in sections 2.4.3 and 2.4.4). In this section, the study's main findings are summarised in relation to the context, cognitive, scoring, criterion-related validity and the test-taker elements of the framework. It should be noted that the scope of the study did not allow for the collection of consequential validity evidence.

The findings from the study serve as complementary sources of validity evidence and are drawn upon to evaluate the extent to which the interpretations made on the basis of test scores are well grounded and not reflective of the unwarranted effects of aspects of the test which are irrelevant to the construct being measured.

In terms of **context validity** evidence, the different topics within each task type were shown to have statistically distinct difficulty levels, meaning that topics designed to be equal can elicit responses which are measurably different. However, these measurable differences failed to reach ‘practical significance’ (Fulcher, 2003). In other words, these are differences that ‘do not matter’ (Dorans and Feigenbaum, 1994), as the maximum difference between the easiest and most difficult topics consistently fell below the minimum speaking ability necessary to move across adjacent bands and across criteria. It can therefore be argued that the topics used in the study, in spite of differences in difficulty measures, can be considered ‘practically equivalent’ and unlikely to exert an influence on the scores at any point across the scale. The only isolated case where this maximum difference could potentially have a practical impact is for the TD criterion (and limited only to specific categories on the scale). The topics (within each task type) were also compared in respect to the functions they elicit. The results revealed both differences and similarities in the range of elicited functions which can be interpreted to imply that different topics can potentially tap into different features of the underlying speaking construct. However, a more in-depth examination of the functions is necessary for making conclusive remarks regarding topic comparability. The results are suggestive of a potential lack of topic comparability in terms of the elicited functions though as discussed earlier, any impact of topics on the language produced by test takers have been shown to not have a large effect on scores.

A more serious threat to context validity of the topics was observed in the systematic, consistent and measurably different ways in which test takers’ level of background knowledge of topics exerted an influence on performance. The qualitative analyses of the

content of spoken performances also revealed a complex interaction between topics and various test-taker characteristics (e.g. cultural background, personal interest, experiential characteristics and affective schemata). This evidence is suggestive of the systematic influence of a task-induced construct-irrelevant variance on performance and can weaken the argument for the context validity of the task. Nevertheless, similar to the results of topics, differences in background knowledge levels (despite statistical significance) failed to have a practical impact on scores (with the exception of the TD criterion), as the maximum level of difference between different background knowledge levels consistently fell below the minimum speaking ability required to move across adjacent band scores and across different criteria.

In terms of **cognitive validity**, the differences in topic difficulty measures can be assumed to reflect test developers' intentions of varying the levels of cognitive demand put on the test takers as a function of topic familiarity and abstractedness of tasks. However, the sequencing of topics did not always follow the intended progression in cognitive demand from easy to difficult (or from familiar to abstract). According to the study's results, the nature of background knowledge is highly test-taker dependent which makes it difficult, if not impossible to make *a priori* assumptions of familiarity or unfamiliarity. This might lead to an inadequate categorisation of tasks in terms of the level of cognitive demand.

In terms of **scoring validity**, the results of MFRM indicated that the raters in the study, though not trained, exhibited high levels of consistency within themselves. The raters exercised significantly different severity levels. However, by adopting an MFRM approach, rater severity was directly parameterized in the model and the raw scores of examinees were adjusted for rater differences. There was also very little evidence of other

systematic rater tendencies in the measurement system. An in-depth examination of the impact of the rater, topic, and background knowledge of topic on scores revealed that, despite significant differences in the element measures of each facet, the differences were too small to have a practical effect on scores. It can therefore be inferred that scores on the test predominantly reflect the underlying speaking construct that the test is designed to tap into and that the speaking scores are not unduly affected by construct-irrelevant factors.

Evidence for the **criterion-related** validity of the speaking test comes from the construction of two parallel forms of the test from a combination of the easiest vs. the most difficult topics in the study and a comparison of resulting measures which showed only negligible differences across the measures of the two forms. The strong positive correlation between the examinee measures from the speaking tests and the C-test results (measure of general language proficiency) can be shown as further evidence of criterion-related validity.

The last element to consider is the **test taker** in the validation framework (Weir, 2005). The findings have shown that some of the test topics in this international speaking test were not necessarily relevant to the Iranian test takers in the study as a result of their cultural background, thus suggesting topic bias bringing the local validity of some of the speaking task topics into question (O'Sullivan, 2011). On the other hand, the qualitative analyses of the content of test-taker performances suggested that other test taker-related variables such as previous experiences, affective schemata, religious background, personal interest, etc. were also in interaction with the topics. It was also demonstrated empirically that different levels of background knowledge can consistently, systematically and significantly affect performance measures. However, as mentioned earlier, with the

exception of the TD criterion, these effects did not reach ‘practical significance’. Results of a bias analysis also indicated that when background knowledge levels are low, test takers from a low proficiency group were at a disadvantage compared to higher proficiency candidates. Topic unfamiliarity was associated, at times, with topic abandonment, disengagement from the questions, and fewer opportunities for test takers to elaborate on topics or questions. Moreover, lack of background knowledge was associated with negative affective influence on test takers, particularly lower proficiency individuals. Taken together, these findings raise some fairness and validity concerns which are associated with topics of the test and should not be neglected.

Broadly speaking, the results of the study lend strong support to the topic validity of the speaking test under examination, as the quantitative findings of the study have consistently shown that large differences in topic measures and test takers’ background knowledge of topics fail to have a practical impact on test scores. Both quantitative and qualitative results of the study indicate that the random assignment of topics and test takers’ level of background knowledge of topics may introduce some bias to the test. In evaluating bias, McNamara and Roever (2006:82) refer to ‘construct-irrelevant variance that distorts the test results and therefore makes conclusions based on scores less valid’. The findings of the study have rejected the absence of bias but have nevertheless shown that (any) bias is not large enough to ‘distort’ test results dramatically or contest the plausibility of interpretations on the basis of test scores (Kane, 2001). Put differently, it is speaking ability, as operationalised in the test, that is the principal determinant of test scores.

This is not to say that the evidence brought forward in respect to the influence of topic and BK on test takers, raters, features of performance and raters’ decision-making should be

ignored. In the words of Messick (1989:13), ‘validity is a matter of degree, not all or none’. The validity issues and concerns raised in this study need to be taken into account and steps should be taken to increase the topic validity of tests. This brings us to the implications of the study’s findings.

7.3 Implications

The study’s main implication is that in second language oral performance assessment contexts, the topic of the performance tasks and the level of background knowledge that test takers bring to the topics are unlikely to have an effect on the final scores assigned to test takers. While there might be some evidence of construct-irrelevant variance and test bias attributable to these two variables, their impact (at least at the score level) is not large enough to distort test results. Attempts should be made to reduce (any) negative impact of topics and background knowledge of topics.

The findings of this study hold further methodological, theoretical and practical implications for speaking performance assessment, which are outlined as follows.

7.3.1 Methodological Implications

The study has two important methodological implications. Firstly, the study empirically demonstrated the various advantages of using the Rasch family of models in examining the quality of the measurement instruments used in the study (Tennant and Conaghan, 2007). In using Multi-Faceted Rasch Measurement (Linacre, 1989) the study illustrated how the different facets of the assessment context can be examined independently and/or in relation to each other and on the same frame of reference. The study’s findings attest to the usability of the Rasch family of models in second language performance assessment

contexts and in instrument design and validation. Secondly, the study demonstrated how the procedures proposed by Marais and Andrich (2011) in detecting ‘a common rater halo effect’ in their study with simulated data was applied to the empirical score data in the study. The results, although inconclusive, pointed to the potential inadequacies of MFRM in detecting such a halo effect (Marais and Andrich, 2011). Thirdly, the study illustrated how the use of a mixed-methods strategy of inquiry (Creswell, 2003; Creswell and Plano-Clark, 2007) can deepen our understanding of the phenomenon under examination. In bringing the quantitative analyses of score data, functions checklist, test-taker questionnaire responses together with the qualitative analyses of rater interviews and the content of test taker performances, the mixed-methods approach to data collection, analysis and interpretation illustrated how the different strands of research can complement one another, provide different types of evidence and shed light on inconsistencies in findings.

7.3.2 Theoretical Implications

On the basis of the findings, the theoretical implications of the study are advanced in relation to facets of speaking performance assessment, test performance models, conceptualisations of task difficulty and speaking test construction definition.

The study has contributed to the body of scholarly work exploring the impact of various facets of the assessment context (examinees, raters, topics, background knowledge of topics, rating criteria) on performance scores. In particular, the impact of the topic and background knowledge of topic facets on performance scores was investigated. Results suggested a small yet statistically significant impact of both variables on performance. These findings were in line with Fulcher (2003) and Fulcher Marquez-Reiter’s (2003)

assertions in respect to the lack of a practical effect of the task condition of topic on performance. Differences in topics and levels of background knowledge were shown to be not large enough to ‘matter’ (Dorans and Feigenbaum, 1994).

The results of the MFRM analyses and the qualitative analyses of data revealed the complexities involved in performance assessment. For example, the raters in this study exhibited different severity and leniency levels in rating, they differentially interacted with the test takers’ proficiency levels and expressed different views in regard to their decision-making process. An examination of the content of test takers’ performances and raters’ interview transcripts revealed the intricate ways in which test takers engaged with topics, their different approaches to dealing with problematic questions and different affective reactions to topics. The observed interactions between various elements of the assessment context in this study lend support to theoretical and psychometric models of language performance and validation that conceptualise speaking ability and performance not as a static entity but one which is in interaction with its surrounding context (e.g. McNamara, 1996; Weir, 2005; O’Sullivan and Weir, 2011; Eckes, 2009).

Another important finding of the study pertained to the conceptualisation of the notion of task or topic ‘difficulty’. The MFRM results of the study illustrated the variations in topic difficulty indices once test takers’ background knowledge of topics was directly parameterized in the model. Moreover, the qualitative analyses of the study revealed the different ways in which test takers’ background characteristics engaged with the topics of the tasks. The results of the study lends strong support to Bachman’s (2002) argument in which he contends that ‘difficulty’ is not an inherent characteristic of the task but rather one that lies in the interaction between the task and characteristics of the test takers and

the assessment context. The findings of the current study have demonstrated that notions such as ‘difficulty’ or ‘familiarity’ cannot be necessarily established *a priori*, as the question is not ‘how difficult?’ but rather, ‘difficult for whom?’. By the same token, it is argued that the notion of ‘parallel or equivalent’ tasks or topics may be misleading. These arguments can be drawn upon to explain why research endeavours in predicting task difficulty (Skehan, 1998; Norris et al., 2002) on the basis of task characteristics alone may have failed to reach conclusive findings or may have shown inconsistencies. In line with Bachman’s (2002:369) it is therefore recommended that a distinction be made between characteristics of the task and the test takers and to conceptualise difficulty as interactions between the two.

The final theoretical implication of the study pertains to the definition of the speaking construct underlying the test. As previously explained, the topic development (TD) criterion did not explicitly form part of the IELTS rating scale and was included as means of isolating the effects of topic and BK of topic. The results of the MFRM analyses showed that the criterion contributed uniquely and independently to explaining variance in speaking scores. Moreover, raters remarked on the importance of the TD criterion in capturing the extent of candidates’ communicative success in addressing a topic. These results are in line with the empirical research carried out by Sato (2012:237) in which a similar criterion of ‘content elaboration/development’ was identified ‘as an additional dimension that is highly relevant to language proficiency but is not fully delineated by existing communicative language ability models’. While the results of the current study are too limited to make any conclusive remarks regarding the role of TD, it is argued that such a criterion can, in theory, serve as an additional component of the speaking construct

in a similar way that the IELTS writing scales include a *task response* criterion and the TOEFL iBT speaking rubrics include a *task fulfillment and topic development* criterion, the latter of which was used in the current study.

The inclusion of such a criterion in the definition of the construct can have both advantages and disadvantages. A possible benefit of a TD criterion (or a similar content-based criterion) is an increased emphasis on the communication of meaning rather than solely focusing on the linguistic quality of performance (Sato, 2012; McNamara, 1996). This can also lead to positive washback in the classroom by aligning assessment criteria with communicative language teaching practices. On the other hand, the inclusion of the criterion could potentially increase the cognitive demand on raters, as they would have to assign scores on an additional criterion and for each individual task. Moreover, in light of the study's findings, the inclusion of such a criterion would also have a potential disadvantage of inflating the impact of background knowledge on performance and would therefore necessitate mechanisms for minimising the (negative) impact of BK (to be discussed in the next section) and/or providing separate weightings for the different criteria. As mentioned before, if a dimension such as TD is not included in the definition of the construct, a validity threat of potential construct underrepresentation might be introduced to a test. These different arguments resonate with Davies' (2008) point in relation to the balancing act between theoretical stances in defining constructs and the practical demands of large scale assessment. Any decision regarding the inclusion of a content-oriented criterion would have to be made in light of both theoretical and practical considerations.

7.3.3 Practical Implications and Recommendations

The study's findings hold practical implications for speaking performance assessment.

Firstly, the findings related to the interaction of topic of the tasks and the test takers' background knowledge of topics suggest that it is very difficult, if not impossible, to establish topic difficulty and/or familiarity *a priori*. As Kasper and Ross (2007: 2065) point out, sensitivity to topics 'is no inherent attribute of that topic but something that participants orient to through their interactional conduct and thereby construct in the first place'. For large-scale standardised tests with an international candidature, the common practice for ensuring that topics of a test are general, comparable, equally familiar or abstract involves, in the most part, the piloting of topics with representative samples of test takers and statistical analyses of scores for potential bias. While these are important endeavours, the study's findings illustrate that these procedures may be inadequate, as the highly individual way in which test takers interact with topics circumvents the possibility of making generalisations regarding topic familiarity/difficulty. What is suggested is to consider shifting some of these efforts to focus on minimising any negative impact of background knowledge, ensuring that mechanisms are in place for dealing with problematic topics and/or the lack of background knowledge and maximizing the opportunities to speak. Some possibilities include:

- Implementing a choice mechanism

Providing test takers with a selection of topics to choose from seems to be the most straightforward approach in addressing many of the problems raised in the study. In line with Jennings et al. (1999), a choice mechanism is likely to reflect the complexity of interactions between different test taker variables and the available topics. By

giving test takers a choice, different objectives can be simultaneously accomplished. For example, the probability of conflict in terms of topic mismatch with test taker background knowledge is reduced. Secondly, by giving test takers a choice, they are also given some power and autonomy (also recommended by Kenyon and Malone, 2010) which can help reduce the asymmetrical power relations between examiners and test takers in the context of a language proficiency interview. Thirdly, the negative impact of unfamiliar or problematic topic/question is likely to be reduced when a choice is offered. The findings in the study have already shown that the topic and BK of topics have no practical impact on performance scores. A choice mechanism would therefore be in place to facilitate speech generation and ensure that test takers are presented with equal opportunities for generating speech without facing topic-related problems.

- Leveling the playing field

Another possible approach for reducing the negative impact of background knowledge is to provide test takers with the necessary information to respond to tasks for example in the form of information-based prompts. Another way of reducing the topic effect is to move away from a completely topic-driven test format to one which includes a wider variety of speaking tasks which do not necessarily require test takers to draw on their background knowledge (e.g. integrated tasks⁵⁰).

- Flexibility in the use of the examiner script

⁵⁰ This is not to say that BK does not exert an influence on performance in integrated tasks but that the influence may be reduced as evidenced in Jennings et al. (1999) and Lee and Anderson (2007).

It was suggested earlier that examiners might feel constrained by the examiner script in not allowing them to deal with the occasional cases where topic unfamiliarity may hinder performance. In line with Linacre (2011a) who argues for a view of raters as ‘independent experts’ and not ‘scoring machines’, a more flexible script may equally allow examiners to act as experts rather than as test delivery/scoring machines so that they would be able to deal with problematic performances.

The second practical implication pertains to the TD criterion. If a TD criterion is included in a speaking test within a multi-task test format, then scores have to be ideally assigned at a task level (and not at the test level) as ‘a single score for performance on a number of tasks does not offer a true reflection of a candidate’s true ability’ (O’Sullivan and Nakatsuhara, 2011: 182). Such a scoring system however, can increase in the cognitive demand posed on examiners and/or raters. Decisions would therefore have to be made regarding the feasibility of marking the test by task within the practical restraints of the assessment context.

The absence of a practical effect of topic and background knowledge of topic on performance scores has an important implication for test takers. As evidenced in the questionnaire responses, test takers believe that the topics assigned to them and their ideas about a topic can have an affect on their scores. This might be a source of anxiety (Bachman and Palmer, 1996; Huang, 2010), particularly in live exam conditions. However, the study’s findings suggest that test takers should not be overly concerned with the topics they might be assigned. They should instead focus on developing their responses. As Seedhouse and Harris (2011) remarked, answering questions do not necessarily coincide with developing and elaborating on a question. Raising awareness of

the importance of question/topic development can be advantageous to test takers, as their final scores would better capture their speaking abilities.

7.4 Limitations

While the present study has contributed to an understanding of the role of topic and background knowledge of topics in second language performance assessment contexts, there are several limitations which have to be acknowledged with regard to the scope and design of the study. These limitations are listed below.

1. The most important limitation of the study is its context which largely pertains to a particular test of speaking with specific features (e.g. test format, topics, rating scales). This can in turn reduce the extent to which findings are applicable to other test settings. Furthermore, the data from the study were not derived from live exam conditions, and the experimental settings of the study diverged from the operational exam settings (e.g. using non-IELTS raters, marking by task and not by test, including a TD criterion), which can limit the direct application of the study's findings to the specific test that was examined.
2. The study's analyses and conclusions pivoted largely on the notion of 'practical significance' (Fulcher, 2003) or 'differences that matter' (Dorans and Feigenbaum, 1994) in examining the contribution of differences in topic difficulty and levels of background knowledge on performance scores. This 'indifference threshold' (Dorans and Liu, 2009:13) is defined in terms of the score units of specific test (e.g. IELTS bands) and as such, is highly exam specific. For this reason, it is not possible to predict

whether lack of practical significance in one speaking test holds true for another exam with different score units.

3. The Iranian participants in the study came from a specific cultural background. It is therefore difficult to predict the extent to which findings from this sample of test takers are applicable to an international test taking population.

4. The speaking tests were administered to participants in an unofficial context and while care was taken to simulate live exam conditions as closely as possible, some variations in test-taker performances may be expected across simulated and live conditions. In particular, the anxiety associated with high-stakes tests may inflate the impact of topic and (lack of) background knowledge of topic on performance.

5. Due to practical and economic constraints, only four raters participated in the study which makes any generalizations on the basis of findings related to raters impossible. These include findings related to the diagnosis of a common rater halo effect (Marais and Andrich, 2011) and findings from a bias analyses which showed how raters' severity/leniency interacted with test takers' general language proficiency in two opposite directions. Given the small sample size, it was not possible to determine whether observations were representative of a trend or were simply isolated cases.

6. The analyses of the functions in the study focused only on the comparison of functions across topics (within task types). A more in-depth analysis can include similar comparisons while also controlling for proficiency level and background knowledge in order to better examine the sources of variations in observed functions.

7. The quantitative analyses of the study involved the inclusion and removal of the Topic Development criterion in the analyses in order to examine the extent to which topic and background knowledge of topic affected performance scores on the remaining criteria (FC, LR, GA, P) once TD was eliminated. However, this procedure was performed mathematically and not experimentally. As a result, it is not possible to predict the influence of topic and BK on the scores of the IELTS criteria had the TD not been included in the scales in the first place.

The above limitations can reduce the generalisability of the study's findings. Notwithstanding these shortcomings, the study adopted a rigorous approach to examining the effects of the variables of interest and systematically controlled for the influence of factors which may have confounded the study's results, en route to providing empirical evidence for addressing important validity concerns in language proficiency interviews.

7.5 Directions for Future Research

In light of the findings of the study and the limitations outlined above, this section proposes several avenues for future research.

Firstly, in order to improve the generalisability of the study, it would be important to replicate the present study with a larger representative sample of the international test-taking population and using a larger pool of raters and topics. While the current study predominantly focused on the test taker-related variable of background knowledge of topic, it would also be useful to include other test taker characteristics such as education, age, gender, level of test-related anxiety and personality traits amongst others (O'Sullivan, 2000b; Nakatsuhara, 2009; Huang, 2010) in order to examine the interactions of these

variables with the topic of tasks. The study can also be replicated in other speaking test contexts and formats (paired and/or group test formats, semi-direct and automated tests), with other task types (e.g. integrated tasks) and with different scales so that findings would no longer be test-specific.

Secondly, in light of preliminary findings pertaining to the TD criterion and in line with Sato (2012:226), it is important to evaluate the role of a non-linguistic and content-oriented criterion within a speaking assessment context and to explore the contribution of test takers' content of performance to 'communicative effectiveness'. The absence of such a criterion in a test may introduce the validity threat of construct underrepresentation. Empirical research into the role of such a criterion within performance assessment contexts is therefore necessary. The study's findings in relation to the potentially large influence of background knowledge on the scores of a content-oriented criterion should also be borne in mind and addressed in any such research endeavors.

Thirdly, it is important to conduct empirical research which compares the effects of topic and BK on scores of linguistically-oriented criteria, when marked under two different conditions: with and without a content-oriented criterion in the rating scales. In doing so, it would be possible to detect whether topic and background knowledge of topic have a smaller or larger impact on linguistically-oriented criteria when raters' attention is not explicitly focused on a content-oriented criterion.

Lastly, further empirical research into the diagnosis of a 'common rater halo effect' (Marais and Andrich, 2011) is recommended. The preliminary results of the study suggested that the common structure of the study's rating scales might have induced a common rater halo effect. However, it was not feasible to determine whether the observed

effect reflected participants' actual performances or the common structure of the rating scale (or both). Given the prevalent use of rating scales in performance assessment contexts and the measurement error which can be potentially introduced to the scores as a result of a halo effect common to all raters, further empirical research in this area is warranted.

7.6 Concluding Remarks

The aim of the present study was to evaluate, from a test validity perspective, the extent to which topic of tasks and the level of background knowledge that test takers bring to the topics have an effect on performance scores in language proficiency interviews. Broadly speaking and within the specific context of the oral test under examination, the results of the study have lent strong support to the topic validity of test, as differences in topics and test takers' level of background knowledge of topics failed to have a practical effect on scores.

The study also identified potential sources of bias and construct-irrelevant variance, linked to the interaction between characteristics of the test takers and topics of the test. Some topics were shown to lack 'local' validity for the majority of test takers in the study while some other topics were not necessarily relevant to specific individuals. Problematic topics and lack of background knowledge about topics appeared to have a negative affective influence on test takers. The variables also exerted an influence on features of performance from the perspective of the raters. Nevertheless, these factors did not exert a large enough influence on scores to distort the interpretations made on the basis of test

scores about test takers' speaking abilities. Put differently, it was the underlying construct of speaking ability that was the major determinant of speaking scores.

Lastly, it was McNamara (2000) who referred to a test score as a 'deceptively' simple number. The present study has revealed some of the many complexities involved in performance assessment and in determining what a single score represents. This thesis aimed to disentangle some of these complexities by focusing on specific aspects of speaking tests, namely topics and test takers' background knowledge of topics. The results serve as simply one piece of the validation puzzle. It is only hoped that the findings are a step in the right direction towards a better understanding of that 'complex web of interaction' (McNamara, 1996) which is speaking performance assessment.

References

- Adobe Systems (2003-2004) *Adobe Audition 1.5*. CA, USA
- Alderson, J.C. and Urquhart, A.H. (1985) The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing*, 2 (2), pp. 192-204.
- Alderson, J.C. and Urquhart, A.H. (1983) The effect of student background discipline on comprehension: A pilot study. In *Current developments in Language Testing*, eds. A. Hughes and D. Porter, London: Academic Press, pp. 121-127.
- American Educational Research Association (1999) *Standards for Educational and Psychological Testing*, American Psychological Association and National Council on Measurement in Education. Washington DC: Author.
- Anastasi, A. (1988) *Psychological testing*. 6th ed. New York: Macmillan.
- Andrich, D. (2005) Georg Rasch: Mathematician and statistician. *Encyclopedia of Social Measurement*, 3, pp. 299-306.
- Andrich, D. (2004) Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical care*, 42 (1), pp. I-7-I-16.
- Andrich, D. (1998) Thresholds, steps, and rating scale conceptualization. *Rasch Measurement: Transactions of the Rasch Measurement*, 12 (3), pp. 648-649.
- Andrich, D. (1985) A latent trait model for items with response dependencies: Implications for test construction and analysis. In *Test Design*, ed. S.E. Emberton, New York: Academic Press, pp. 245-275.
- Andrich, D. (1982) An index of person separation in latent trait theory, the traditional KR. 20 index, and the Guttman scale response pattern. *Education Research and Perspectives*, 9 (1), pp. 95-104.
- Andrich, D. (1978) Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, 2 (4), pp. 581-594.
- Andrich, D., De Jong, J. and Sheridan, B.E. (1997) Diagnostic opportunities with the Rasch model for ordered response categories. *Applications of latent trait and latent class models in the social sciences*, , pp. 59-70.
- Andrich, D. and Hagquist, C. (2012) Real and Artificial Differential Item Functioning. *Journal of Educational and Behavioral Statistics*, 37, pp. 387-416.
- Andrich, D. and Marais, I. (2011) *Advance course in Rasch measurement of modern test theory (course material)*. University of Western Australia.
- Andrich, D., Lyne, A., Sheridan, B. and Luo, G. (2010) *RUMM 2030*. Perth: RUMM Laboratory

- Andrich, D. and Marais, I. (2010) *Introduction to Rasch measurement of modern test theory (course material)*. University of Western Australia.
- Babaii, E. and Ansary, H. (2001) The C-test: a valid operationalization of reduced redundancy principle? *System*, 29 (2), pp. 209-219.
- Bachman, L.F. (2011) *How do different language frameworks impact language assessment practice?* Paper presented at ALTE 4th International Conference in Krakow (July, 2011)
- Bachman, L.F. (2004) *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.
- Bachman, L.F. (2002) Some reflections on task-based language performance assessment. *Language testing*, 19 (4), pp. 453-476.
- Bachman, L.F. (1990) *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.
- Bachman, L.F., Lynch, B.K. and Mason, M. (1995) Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, 12 (2), pp. 238-257.
- Bachman, L.F. and Palmer, A.S. (1996) *Language testing in practice*. Oxford: Oxford University Press.
- Baker, B.A. (2012) Individual differences in rater decision-making style: an exploratory mixed-methods study. *Language Assessment Quarterly*, 9 (3), pp. 225-248.
- Barkaoui, K. (2010) Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, 7 (1), pp. 54-74.
- Barkaoui, K. (2007) Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, 12 (2), pp. 86-107.
- X. Bei. (2010) *The effects of topic familiarity and strategic planning in topic-based task performance at different proficiency levels*. Unpublished PhD thesis. Chinese University of Hong Kong, China.
- Bond, T.G. and Fox, C.M. (2007) *Applying the Rasch model: Fundamental measurement in the human sciences*. 2nd ed. New Jersey: Lawrence Erlbaum Associates, Inc.
- Brooks, L. (2003) Converting an observation checklist for use with the IELTS speaking test. *Cambridge ESOL Research Notes*, 11, pp. 20-21.
- Brooks, L. (2002) Report on functions observed in the old IELTS Speaking Test versus those in the revised Speaking Test, Internal Cambridge ESOL Report, Cambridge.

- Brown, A. (2007) An investigation of the rating process in the IELTS oral interview In *IELTS collected papers: research in speaking and writing assessment.*, eds. L.B. Taylor and P. Falvey, Cambridge: Cambridge University Press, pp. 98-139.
- Brown, A. (2006) An examination of rating process in the revised IELTS Speaking Test. In *IELTS Research Reports Volume 6.*, eds. P. McGovern and S. Walsh, IELTS Australia and British Council, pp. 41-70.
- Brown, A. and Hill, K. (1998) Interviewer style and candidate performance in the IELTS oral interview. In *IELTS Research Reports Volume 1.*, ed. S.Woods Sydney: ELICOS.
- Brown, A., Iwashita, N. and McNamara, T. (2005) *An examination of rater orientations and test-taker performance on English-for-Academic-Purposes speaking tasks (TOEFL Monograph No. 29)*. Princeton, NJ: Educational Testing Service.
- Brown, A. and Taylor, L. (2006) A worldwide survey of examiners' views and experience of the revised IELTS Speaking test. *Cambridge ESOL Research Notes*, 26, pp. 14-18.
- Brown, G., Anderson, A., Shillcock, R. and Yule, G. (1984) *Teaching talk: strategies for production and assessment*. Cambridge: Cambridge University Press.
- Canale, M. (1987) The measurement of communicative competence. *Annual Review of Applied Linguistics*, 8, pp. 67-84.
- Carrell, P.L. (1981) Culture-specific schemata in L2 comprehension. *Selected Papers from the 9th Illinois TESOL/BE Annual Convention and the first Midwest TESOL Conference*, eds. R.A. Orem & J.F. Haskell, Chicago: Illinois TESOL/BE.
- Carrell, P.L. and Wise, T.E. (1998) The relationship between prior knowledge and topic interest in second language reading. *Studies in Second Language Acquisition*, 20 (3), pp. 285-309.
- Chen, Q. and Donin, J. (1997) Discourse Processing of First and Second Language Biology Texts: Effects of Language Proficiency and Domain-Specific Knowledge. *The Modern Language Journal*, 81 (2), pp. 209-227.
- Chiang, C.S. and Dunkel, P. (1992) The effect of speech modification, prior knowledge and listening proficiency on EFL lecture learning. *TESOL Quarterly*, 26 (2), pp. 345-374.
- Cizek, G.J. (2011) Reconceptualizing validity and the place of consequences. *Annual Meeting of the National Council on Measurement in Education*. New Orleans
- Clapham, C. (2000) Assessment for academic purposes: where next? *System*, 28 (4), pp. 511-521.
- Clapham, C. (1996) *The development of IELTS: a study of the effect of background knowledge on reading comprehension*. Studies in Language Testing 4, Cambridge: Cambridge University Press.
- Cohen, A., Segal, M., Weiss, R., (1984) The C-test in Hebrew. *Language Testing*, 1, pp. 221-225.

- Constable, E. and Andrich, D. (1984) *Inter-Judge reliability: is complete agreement among judges the ideal?*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans
- Creswell, J.W. (2003) *Research Design: Qualitative, Quantitative and Mixed Methods Approaches*. 2nd ed. California, the United States of America: Sage Publications Inc.
- Creswell, J.W. and Plano- Clark, V.L. (2007) *Designing and conducting mixed methods research*. Wiley Online Library.
- Cronbach, L.J. (1990) *Essentials of psychological testing*. 5th ed. New York: Harper and Row.
- Cronbach, L.J. (1988) Five perspectives on validity argument. In *Test validity*., ed. H. Wainer, Hillsdale, NJ: Erlbaum, pp. 3-17.
- Cronbach, L.J. (1951) Coefficient alpha and the internal structure of tests. *Psychometrika*, 16 (3), pp. 297-334.
- Cumming, A., Kantor, R. and Powers, D.E. (2002) Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, 86 (1), pp. 67-96.
- Davies, A. (2008) *Assessing academic English: testing English proficiency, 1950-89 - the IELTS solution*. Studies in Language Testing 23, Cambridge: UCLES/Cambridge University Press.
- Dorans, N.J. and Liu, J. (2009) *Score equity assessment: development of a prototype analysis using SAT Mathematics test data across several administrations*. ETS Research Report No. RR-09-08. Princeton, NJ: Educational Testing Service.
- Dorans, N. and Feigenbaum, M. (1994) Equating issues engendered by changes to the SAT and PSAT/NMSQT. *Technical issues related to the introduction of the new SAT and PSAT/NMSQT*, pp. 91-122.
- Douglas, D. and Selinker, L. (1992) Analyzing oral proficiency test performance in general and specific purpose contexts. *System*, 20 (3), pp. 317-328.
- Eckes, T. (2010) The TestDaF implementation of the SOPI: design, analysis, and evaluation of a semi-direct speaking test. *Proceedings of the Computer-based Assessment (CBA) of Foreign Language Speaking Skills*, ed. L. Araujo, European Union, pp. 63-83.
- Eckes, T. (2009) Many-facet Rasch measurement. In *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*., ed. S. Takala, Strasbourg, France: Council of Europe: Language Policy Division. Retrieved from www.coe.int/t/dg4/Linguistic/CEF-refSupp-SectionH.pdf, pp. (Section H).
- Eckes, T. and Grotjahn, R. (2006) A closer look at the construct validity of C-tests. *Language Testing*, 23 (3), pp. 290.

- Edgeworth, F.Y. (1890) The element of chance in competitive examinations. *Journal of the Royal Statistical Society*, 53 (4), pp. 644-663.
- Educational Testing Service (2004) *TOEFL iBT Next Generation TOEFL Test (Independent Speaking Rubrics: Topic Development)*. Available from: <www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf2004> (Accessed December, 2012).
- Ellis, R. and Barkhuizen, G.P. (2005) *Analysing learner language*. Oxford University Press Oxford.
- English-Online (n.d.) *The Nobel Prize [Online]*. Available from: <<http://www.english-online.at/society/nobel-prize/nobel-prize.htm>> (Accessed May, 17, 2010).
- English-Online (n.d.) *Pablo Picasso [Online]*. Available from: <<http://www.english-online.at/art-architecture/pablo-picasso/picasso-master-of-cubism.htm>> (Accessed May, 17, 2010).
- English-Online (n.d.) *Street Art- New Culture of the Cities [Online]*. Available from: <<http://www.english-online.at/art-architecture/street-art/graffiti-street-artists.htm>> (Accessed May, 17, 2010).
- Fanselow, J.F. (1977) Beyond Rashomon: conceptualizing and describing the teaching act. *TESOL Quarterly*, 11 (1), pp. 17-39.
- Farrokhi, F. and Esfandiari, R. (2011) A Many-facet Rasch model to detect halo effect in three types of raters. *Theory and Practice in Language Studies*, 1 (11), pp. 1531-1540.
- Feldmann, U. and Stemmer, B. (1987) Thin_____ aloud a_____ retrospective da_____ in c-te_____ taking:
diffe_____ languages diffe_____ learners -sa_____ approaches? In *Introspection in Second Language Research.*, eds. C. Faerch and C. Kasper, Clevedon: Multilingual Matters, pp. 251-267.
- Field, A. (2010) *Discovering statistics using SPSS (and sex and drugs and rock'n'roll)*. 3rd ed. London: Sage Publications Inc.
- Field, J. (2011) Cognitive validity. In *Examining speaking: research and practice in assessing second language speaking.*, ed. L. Taylor, Studies in Language Testing 30, Cambridge: UCLES/Cambridge University Press, pp. 65-111.
- Flesch, R. (1948) A new readability yardstick. *Journal of applied psychology*, 32 (3), pp. 221-233.
- Flucher, G. and Marquez Reiter, R. (2003) Task difficulty in speaking tests. *Language Testing*, 20 (3), pp. 321-344.
- Foucault, M. (1975) *Discipline and Punish: the birth of the prison*. New York: Vintage.
- Fulcher, G. (2003) *Testing second language speaking*. Harlow: Pearson Education Ltd.

- G. Fulcher. (1993) *The construction and validation of rating scales for oral tests in English as a foreign language*. Unpublished PhD thesis. University of Lancaster, UK.
- Fulcher, N.G. and Davidson, F. (2007) *Language testing and assessment: an advanced resource book*. Oxon: Routledge.
- Galaczi, E. and French, A. (2011) Context validity. In *Examining Speaking*, ed. L. Taylor, Cambridge: Cambridge University Press, pp. 112-170.
- Goetz, J.P. and LeCompte, M.D. (1984) *Ethnography and qualitative design in educational research*. Academic Press Orlando, FL.
- Guttman, L. (1950) The basis for scalogram analysis. *Measurement and prediction*, 4, pp. 60-90.
- Hambleton, R.K. and Rodgers, J. (1995) Item bias review. *Practical Assessment, Research, and Evaluation*, 4 (6). Retrieved January 4, 2013 from <http://PAREonline.net/getvn.asp?v=4&n=6>
- Hastings, A.J. (2002) In defense of C-testing. In *Der C-Test: theoretische Grundlagen und praktische Anwendungen [The C-test: theoretical foundations and practical applications]*, ed. R. Grotjahn, Volum 4 Bochum: AKS-Verlag, pp. 11-25.
- He, L. and Shi, L. (2012) Topical knowledge and ESL writing. *Language Testing*, 29 (3), pp. 443-464.
- H. T. D. Huang. (2010) *Modeling the relationships among topical knowledge, anxiety, and integrated speaking test performance: a structural equation modeling approach*. Unpublished PhD. Thesis. University of Texas.
- Hughes, A. (2003) *Testing for language teachers*. 2nd ed. Cambridge: Cambridge University Press.
- Hughes, A. (1989) *Testing for language teachers*. Cambridge: Cambridge University Press.
- IBM Statistics (2010) *SPSS 19.0*. Chicago, IL
- International English Language Testing System (IELTS) (2007) *Percentile Ranks 2007*. IELTS. [Online]. Available from: <http://www.ielts.org/researchers/analysis_of_test_data.aspx> (Accessed June, 2009).
- International English Language Testing System (IELTS) (n.d.) *What is IELTS?* IELTS. [Online]. Available from: <www.ielts.org/pdf/Speaking%20Band%20descriptors.pdf> (Accessed 15 December, 2012).
- International English Language Testing System (IELTS) (2011) *IELTS guide for educational institutions, governments, professional bodies and commercial organisations*. IELTS. [Online]. Available from: <www.ielts.org/pdf/Guide_for_Institutions_and_Organisations2011.pdf> (Accessed 11 December, 2012).

- International English Language Testing System (IELTS) (2010) *IELTS information for candidates* [Online]. Available from: <www.ielts.org/pdf/Information_for_Candidates_booklet.pdf> (Accessed 15 September, 2012).
- Jamieson, J.M., Eignor, D., Grabe, W. and Kunnan, A. (2008) Frameworks for a new TOEFL. In *Building a validity argument for the Test of English as a Foreign Language.*, eds. C.A. Chapelle M.K. Enright and J.M. Jamieson, New York: Taylor and Francis Group, pp. 55-95.
- Jennings, M., Fox, J., Graves, B. and Shohamy, E. (1999) The test-takers' choice: an investigation of the effect of topic on language-test performance. *Language Testing*, 16 (4), pp. 426-456.
- Jensen, C. and Hansen, C. (1995) The effect of prior knowledge on EAP listening-test performance. *Language Testing*, 12 (1), pp. 99-119.
- Johnson, P. (1982) Effects on reading comprehension of building background knowledge. *TESOL Quarterly*, 16 (4), pp. 503-516.
- Kamimoto, T. (1992) An inquiry into what a C-test measures. *Fukuoka Women's Junior College Studies*, 44, pp. 67-79.
- Kane, M.T. (2001) Current concerns in validity theory. *Journal of Educational Measurement*, 38 (4), pp. 319-342.
- Kane, M.T. (1992) An argument-based approach to validity. *Psychological bulletin*, 112 (3), pp. 527-535.
- Kasper, G. and Ross, S.J. (2007) Multiple questions in oral proficiency interviews. *Journal of Pragmatics*, 39 (11), pp. 2045-2070.
- Kenyon, D. and Malone, M. (2010) Investigating examinee autonomy in a computerized test of oral proficiency. *Proceedings of the Computer-based Assessment (CBA) of Foreign Language Speaking Skills*, ed. L. Araujo, European Union, pp. 1-28.
- Kenyon, D.M. and Malabonga, V. (2001) Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning & Technology*, 5 (2), pp. 60-83.
- N. Khabbazzbashi. (2008) *The effects of topic variance on candidate performance in the IELTS speaking module*. Unpublished Dissertation. University of Oxford.
- Khalifa, H. and Salamoura, A. (2011) Criterion-related validity. In *Examining speaking: research and practice in assessing second language speaking.*, ed. L. Taylor, Studies in Language Testing 30, Cambridge: UCLES/Cambridge University Press, pp. 259-292.
- Khalifa, H. and Weir, C.J. (2009) *Examining reading*. Cambridge: Cambridge University Press.
- Kincaid, J.P. (1975) *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Springfield, Virginia: National Technical Information Service.

- Kirk, R.E. (1996) Practical significance: a concept whose time has come. *Educational and Psychological Measurement*, 56 (5), pp. 746-759.
- Klein-Braley, C. and Raatz, U. (1984) A survey of research on the C-Test. *Language Testing*, 1, pp. 134-146.
- Klein-Braley, C. (1997) C-Tests in the context of reduced redundancy testing: an appraisal. *Language Testing*, 14 (1), pp. 47-84.
- Krekeler, C. (2006) Language for special academic purposes (LSAP) testing: the effect of background knowledge revisited. *Language Testing*, 23 (1), pp. 99-130.
- Kuder, G.F. and Richardson, M.W. (1937) The theory of the estimation of test reliability. *Psychometrika*, 2 (3), pp. 151-160.
- Kuhn, T.S. (1962) *The structure of scientific revolutions*. Chicago: University of Chicago Press.
- Kunnan, A.J. (1995) *Test Taker Characteristics and Test Performance: A Structural Modelling Approach*. Studies in Language Testing 2, Cambridge: UCLES/Cambridge University Press.
- Kurosawa, A. (1950) *Rashōmon*. Japan: Daiei Film Co., Ltd.
- Lado, R. (1961) *Language testing: the construction and use of foreign language tests: a teacher's book*. New York: McGraw-Hill.
- Lee, H.K. and Anderson, C. (2007) Validity and topic generality of a writing performance test. *Language Testing*, 24 (3), pp. 307-330.
- Lee-Ellis, S. (2009) The development and validation of a Korean C-Test using Rasch Analysis. *Language Testing*, 26 (2), pp. 245-274.
- Levelt, W.J.M. (1989) *Speaking : from intention to articulation*. Cambridge, MA: MIT Press.
- Lewkowicz, J. (1997) The integrated testing of a second language. In *Encyclopedia of language and education: language testing and assessment 7.*, eds. C. Clapham and D. Corson, Dordrecht: The Netherlands: Kluwer, pp. 121-130.
- 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), pp. 543-560.
- G. S. Lim. (2009) *Prompt and rater effects in second language writing performance assessment*. Unpublished PhD thesis. University of Michigan.
- Linacre, J.M. (2011b) *Facets computer program for many-facet Rasch measurement*. Beaverton, Oregon: Winsteps
- Linacre, J.M. (2011a) A User's Guide to FACETS Rasch Model Computer Programs- Program Manual 3.68.1 , pp. 1-311.

- Linacre, J.M. (1994) *Many-facet Rasch measurement*. MESA Press, Chicago.
- Linacre, J.M. (1989) *Many-facet Rasch measurement*. Chicago: MESA Press.
- Linn, R.L. and Gronlund, N.E. (2000) *Measurement and Assessment in Teaching*. 8th ed. Upper Saddle River, NJ: Prentice-Hall.
- Long, D.R. (1990) What you don't know can't help you. *Studies in Second Language Acquisition*, 12, pp. 65-80.
- Luce, R.D. and Tukey, J.W. (1964) Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of mathematical psychology*, 1 (1), pp. 1-27.
- Lumley, T. (2002) Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing*, 19 (3), pp. 246-276.
- Lumley, T. and McNamara, T.F. (1995) Rater characteristics and rater bias: implications for training. *Language Testing*, 12 (1), pp. 54-71.
- Lumley, T. and O'Sullivan, B. (2000) *The effect of speaker and topic variables on task-performance in a tape-mediated assessment of speaking*. Paper presented at the 2nd Annual Asian Language Assessment Research Forum, The Hong Kong Polytechnic University.
- Lumley, T. and 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), pp. 415-437.
- Luoma, S. (2004) *Assessing speaking*. Cambridge, UK: Cambridge University Press.
- Madsen, H.S. (1983) *Techniques in testing*. Oxford: Oxford University Press.
- Malik, A.A. (1990) A psycholinguistic analysis of the reading behavior of EFL-proficient readers using culturally familiar and culturally nonfamiliar expository texts. *American Educational Research Journal*, 27 (1), pp. 205-223.
- Marais, I. and Andrich, D. (2011) Diagnosing a common rater halo effect using the polytomous Rasch model. *Journal of Applied Measurement*, 12 (3), pp. 194-211.
- Marais, I. and Andrich, D. (2008) Formalizing dimension and response violations of local dependence in the unidimensional Rasch model. *Journal of Applied Measurement*, 9 (3), pp. 200-215.
- Markham, P. and Latham, M. (1987) The influence of religion-specific background knowledge on the listening comprehension of adult second-language students. *Language Learning*, 37 (2), pp. 157-170.
- Masters, G.N. (1982) A Rasch model for partial credit scoring. *Psychometrika*, 47 (2), pp. 149-174.

- McNamara, T. (2000) *Language testing*. Oxford: Oxford University Press.
- McNamara, T. (1996) *Measuring second language performance*. New York, United States: Addison Wesley Longman Limited.
- McNamara, T. and Roever, C. (2006) *Language testing: the social dimension*. Oxford: Blackwell Publishing Ltd.
- Mehnert, U. (1998) The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition*, 20 (1), pp. 83-108.
- Messick, S. (1996) Validity and washback in language testing. *Language Testing*, 13 (3), pp. 241-256.
- Messick, S. (1989) Validity. In *Educational Measurement*, ed. R.L. Linn, 3rd ed. New York: American Council on Education, Macmillan Publishing Company, pp. 13-103.
- MetaMetrics (2011) *What is a Lexile Measure?*. Available from: <<http://www.lexile.com/about-lexile/lexile-overview/>> (Accessed 06 January, 2011).
- Milanovic, M., Saville, N. and Shuhong, S. (1996) A study of the decision-making behaviour of composition markers. In *Performance Testing, Cognition, and Assessment: Selected Papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem*, eds. M. Milanovic and N. Saville, Studies in Language Testing 3, Cambridge: UCLES/Cambridge University Press, pp. 92-114.
- Myford, C.M. and Wolfe, E.W. (2004) Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part II. *Journal of Applied Measurement*, 5 (2), pp. 189-227.
- Myford, C.M. and Wolfe, E.W. (2003) Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I. *Journal of Applied Measurement*, 4 (4), pp. 386-422.
- Myford, C.M. and Wolfe, E.W. (2000) *Monitoring sources of variability within the test of spoken English assessment system (Research Report)*. Princeton: Educational Testing Service.
- Nakatsuhara, F. (2011) Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing*, 28 (4), pp. 483-508.
- F. Nakatsuhara. (2009) *Conversational styles in group oral tests: how is the conversation co-constructed?* Unpublished PhD thesis. University of Essex.
- Norris, J., Brown, J.D., Hudson, T. and Yoshioka, J. (1998) *Designing second language performance assessments*. Technical Report No. 18, Hawaii: University of Hawaii Press.
- Norris, J.M., Brown, J.D., Hudson, T.D. and Bonk, W. (2002) Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing*, 19 (4), pp. 395-418.

- Ortega, L. (1999) Planning and focus on form in L2 oral performance. *Studies in second language acquisition*, 21 (1), pp. 109-148.
- O'Sullivan, B. (ed.)(2011) *Language testing: theories and practices*. Basingstoke: Palgrave MacMillan.
- O'Sullivan, B. and Green, A. (2011) Test taker characteristics. In *Examining speaking: research and practice in assessing second language speaking.*, ed. L. Taylor, Studies in Language Testing 30, Cambridge: UCLES/Cambridge University Press, pp. 36-64.
- O'Sullivan, B. and Nakatsuhara, F. (2011) Quantifying conversational styles in group oral test discourse. In *Language testing: theories and practices.*, ed. B. O'Sullivan, Basingstoke: Palgrave MacMillan, pp. 164-185.
- O'Sullivan, B. and Rignall, M. (2007) Assessing the value of bias analysis feedback to raters for the IELTS Writing Module. In *IELTS collected papers: Research in speaking and writing assessment.*, eds. L. Taylor and P. Falvey, Studies in Language Testing 19, Cambridge: Cambridge University Press, pp. 446-478.
- O'Sullivan, B. and Weir, C.J. (2011) Test development and validation. In *Language testing: theories and practices.*, ed. B. O'Sullivan, Hampshire, UK: Palgrave MacMillan, pp. 13-32.
- O'Sullivan, B., Weir, C.J. and Saville, N. (2002) Using observation checklists to validate speaking-test tasks. *Language Testing*, 19 (1), pp. 33-56.
- O'Sullivan, B. (2002) Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, 19 (3), pp. 277-295.
- O'Sullivan, B. (2000b) Exploring gender and oral proficiency interview performance. *System*, 28 (3), pp. 373-386.
- B. O'Sullivan. (2000a) *Towards a model of performance in oral language testing*. Unpublished PhD Thesis. University of Reading.
- Pallant, J.F. and Tennant, A. (2006) An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology*, 46 (1), pp. 1-18.
- Papajohn, D. (1999) The effect of topic variation in performance testing: the case of the chemistry TEACH test for international teaching assistants. *Language Testing*, 16 (1), pp. 52-81.
- Pollitt, A. and Murray, N.L. (1996) What raters really pay attention to. In *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem.*, eds. M. Milanovic and N. Saville, Studies in Language Testing 3, Cambridge: UCLES/Cambridge University Press, pp. 74-91.
- Raatz, U. and Klein-Braley, C. (2002) Introduction to language testing and C-tests. In *University language testing and the C-test.*, eds. J.A. Coleman R. Grotjahn and U. Raatz, Verlag: AKS, pp. 75-91.

- Rasch, G. (1960) Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. .
- Read, J. (1990) Providing relevant content in an EAP writing test. *English for Specific Purposes*, 9 (2), pp. 109-121.
- Rosenbaum, P.R. (1988) Item bundles. *Psychometrika*, 53, pp. 349-359.
- RUMM Laboratory (2009) *RUMM 2030 manuals*.
- Sato, T. (2012) The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing*, 29 (2), pp. 223-241.
- Schmidt-Rinehart, B.C. (1994) The effects of topic familiarity on second language listening comprehension. *The Modern Language Journal*, 78 (2), pp. 179-189.
- Seedhouse, P. and Harris, A. (2011) Topic development in the IELTS Speaking Test. In *IELTS Research Reports Volume 12.*, ed. J. Osborne, IDP IELTS Australia and British Council, pp. 69-124.
- Shohamy, E. (2001) *The power of tests: A critical perspective on the uses of language tests*. Pearson Education.
- Shohamy, E., Gordon, C.M. and Kraemer, R. (1992) The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, 76 (1), pp. 27-33.
- Singleton, D. and Singleton, E. (2002) The C-Test and L2 acquisition/processing research. In *University language testing and the C-test.*, eds. J.A. Coleman R. Grotjahn and U. Raatz, Bochum: AKS-Verlag, pp. 143-168.
- Skehan, P. (2001) Tasks and language performance assessment. In *Researching pedagogic tasks: second language learning, teaching and testing.*, eds. M. Bygate P. Skehan and M. Swain, London: Longman, pp. 167-185.
- Skehan, P. (1998) *A cognitive approach to language learning*. Oxford: Oxford University Press.
- Skehan, P. (1996) A framework for the implementation of task-based instruction. *Applied linguistics*, 17 (1), pp. 38-62.
- Skehan, P. and Foster, P. (1999) The influence of task structure and processing conditions on narrative retellings. *Language Learning*, 49, pp. 93-120.
- Skehan, P. and Foster, P. (1997) Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research*, 1 (3), pp. 185-211.
- Skehan, P., Xiaoyue, B., Qian, L. and Wang, Z. (2012) The task is not enough: processing approaches to task-based performance. *Language Teaching Research*, 16 (2), pp. 170-187.

- Smith, J. (1989) Topic and variation in ITA oral proficiency: SPEAK and field-specific tests. *English for Specific Purposes*, 8, pp. 155-167.
- J. A. Smith. (1992) *Topic and variation in oral proficiency in international teaching assistants*. PhD thesis. University of Minnesota, DA.
- S. Smith. (2009) *IELTS Examination Preparation among University of Oxford Post-graduate students*. Unpublished Msc. dissertation. University of Oxford.
- Spolsky, B. (1981) Some ethical questions about language testing. In Klein-Braley, C. and Stevenson, D.K., editors, *Practice and problems in language testing 1*, Frankfurt/M.: Lang. In *Practice and problems in language testing.*, eds. C. Klein-Braley and D.K. Stevenson, Frankfurt: Lang.
- Steffensen, M.S., Joag-Dev, C. and Anderson, R.C. (1979) A cross-cultural perspective on reading comprehension. *Reading Research Quarterly*, 15 (1), pp. 10-29.
- Stobart, G. (2008) *The validity of ability tests-case of over-interpretation?* 34th International Association for Educational Assessment (IAEA) Annual Conference in Cambridge (September 2008)
- Strauss, A.L. and Corbin, J.M. (1998) *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. U.S.A: Sage Publications Inc.
- Tarone, E. (1998) Research on interlanguage variation: implications for language testing. In *Interfaces Between Second Language Acquisition and Language Testing Research.*, eds. L. Bachman and A.d. Cohen, UK: Cambridge University Press, pp. 71-89.
- Tarone, E. (1988) *Variation in Interlanguage*. London: Edward Arnold.
- Tashakkori, A. and Teddlie, C. (2002) *Handbook of mixed methods in social & behavioral research*. Sage Publications, Incorporated.
- Taylor, L. (ed.) (2011) *Examining Speaking: Research and practice in assessing second language speaking*. Cambridge: Cambridge University Press.
- Taylor, L. (2011) Introduction. In *Examining speaking: research and practice in assessing second language speaking.*, ed. L. Taylor, *Studies in Language Testing* 30, Cambridge: UCLES/Cambridge University Press, pp. 1-35.
- Taylor, L. (2007b) The impact of the joint-funded research studies on the IELTS Speaking Module. In *IELTS Collected Papers: Research in speaking and writing assessment.*, eds. L. Taylor and P. Falvey, *Studies in Language Testing* 19, Cambridge: Cambridge University Press, pp. 185-194.
- Taylor, L. (2007a) Introduction. In *IELTS Collected Papers: Research in speaking and writing assessment.*, eds. L. Taylor and P. Falvey, *Studies in Language Testing* 19, Cambridge: Cambridge University Press, pp. 1-36.

- Taylor, L. and Taylor, L. (2011) Scoring validity. In *Examining speaking: research and practice in assessing second language speaking.*, ed. L. Taylor, Studies in Language Testing 30, Cambridge: UCLES/Cambridge University Press, pp. 171-233.
- Tedick, D.J. (1990) ESL writing assessment: subject-matter knowledge and its impact on performance. *English for Specific Purposes*, 9 (2), pp. 123-143.
- Tennant, A. and Conaghan, P.G. (2007) The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care & Research*, 57 (8), pp. 1358-1362.
- Time Out (2007) *'Time Out' 1000 books to change your life*. Time-Out Group Ltd.
- University of Cambridge ESOL Examinations (2009) *Cambridge IELTS 7: Examination papers from University of Cambridge ESOL Examinations*. Cambridge: Cambridge University Press.
- University of Cambridge ESOL Examinations (2008) *Cambridge IELTS 6: Examination papers from University of Cambridge ESOL Examinations*. Cambridge: Cambridge University Press.
- University of Cambridge ESOL Examinations (2006) *Cambridge IELTS 5: Examination papers from University of Cambridge ESOL Examinations*. Cambridge: Cambridge University Press.
- University of Cambridge ESOL Examinations (2006) *IELTS Scores Explained (DVD)*.
- University of Cambridge ESOL Examinations (2005) *Cambridge IELTS 4: Examination papers from University of Cambridge ESOL Examinations*. Cambridge: Cambridge University Press.
- University of Cambridge ESOL Examinations (2002) *Cambridge IELTS 3: Examination papers from the University of Cambridge Local Examination Syndicate*. Cambridge: Cambridge University Press.
- Uso-Juan, E. (2006) The compensatory nature of discipline-related knowledge and English-language proficiency in reading English for Academic Purposes. *The Modern Language Journal*, 90 (2), pp. 210-227.
- Van Moere, A. (2010) Automated spoken language testing: test construction and scoring. *Proceedings of the Computer-based Assessment (CBA) of Foreign Language Speaking Skills*, ed. L. Araujo, European Union, pp. 84-99.
- Van Wyke, J. and Andrich, D. (ARC Linkage Report No. 2, 2006) *A typology of polytomously scored mathematics items disclosed by the Rasch model: implications for constructing a continuum of achievement*. Perth, Western Australia: .
- Weigle, S.C. (2004) Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing*, 9 (1), pp. 27-55.
- Weigle, S.C. (1998) Using FACETS to model rater training effects. *Language Testing*, 15 (2), pp. 263-287.

- Weiner, B. (1992) *Human motivation: metaphors, theories, and research*. California: Sage Publications.
- Weir, C., O'Sullivan, B. and Horai, T. (2006) Exploring difficulty in speaking tasks: an intra-task perspective. In *IELTS Research Reports: Volume 6.*, eds. P. McGovern and S. Walsh, United Kingdom, Australia: IELTS Australia and British Council, pp. 119-160.
- Weir, C.J. (2005) *Language testing and validation: an evidence-based approach*. Basingstoke: Palgrave Macmillan.
- Weir, C.J. (1988) *Communicative language testing*. Exeter: Exeter University Press.
- Weir, C.J., Vidakovic, I. and Galaczi, E.D. (in press) *Measured constructs: a history of the constructs underlying Cambridge English examinations 1913-2012*. Cambridge: University of Cambridge ESOL Examinations/Cambridge University Press.
- Weir, C.J. and Wu, J.R.W. (2006) Establishing test form and individual task comparability: a case study of a semi-direct speaking test. *Language Testing*, 23 (2), pp. 167-197.
- Wigglesworth, G. (1997) An investigation of planning time and proficiency level on oral test discourse. *Language Testing*, 14 (1), pp. 85-106.
- Wigglesworth, G. (1993) Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, 10 (3), pp. 305-319.
- Wiseman, C.S. (2012) Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, 17 (3), pp. 150-173.
- Wright, B.D. and Masters, G.N. (2002) Number of person or item strata. *Rasch Measurement Transactions*, 16 (3), pp. 888.
- Yorozuya, R. and Oller Jr, J.W. (1980) Oral proficiency scales: Construct validity and the halo effect. *Language Learning*, 30 (1), pp. 135-153.

Appendices

Appendix A. IELTS 9-Band Scale (IELTS, 2011:12)

9	Expert user	Has fully operational command of the language: appropriate, accurate and fluent with complete understanding.
8	Very good user	Has fully operational command of the language with only occasional unsystematic inaccuracies and inappropriacies. Misunderstandings may occur in unfamiliar situations. Handles complex detailed argumentation well.
7	Good user	Has operational command of the language, though with occasional inaccuracies, inappropriacies and misunderstandings in some situations. Generally handles complex language well and understands detailed reasoning.
6	Competent user	Has generally effective command of the language despite some inaccuracies, inappropriacies and misunderstandings. Can use and understand fairly complex language, particularly in familiar situations.
5	Modest user	Has partial command of the language, coping with overall meaning in most situations, though is likely to make many mistakes. Should be able to handle basic communication in own field.
4	Limited user	Basic competence is limited to familiar situations. Has frequent problems in understanding and expression. Is not able to use complex language.
3	Extremely limited user	Conveys and understands only general meaning in very familiar situations. Frequent breakdowns in communication occur.
2	Intermittent user	No real communication is possible except for the most basic information using isolated words or short formulae in familiar situations and to meet immediate needs. Has great difficulty understanding spoken and written English.
1	Non user	Essentially has no ability to use the language beyond possibly a few isolated words.
0	Did not attempt the test	No assessable information provided.

The IELTS 9-Band Scale (Reproduced from IELTS, 2011:12)

Appendix B. IELTS Speaking Scale (IELTS, n.d.)

IELTS™

SPEAKING: Band Descriptors (public version)

Band	Fluency and coherence	Lexical resource	Grammatical range and accuracy	Pronunciation
9	- speaks fluently with only rare repetition or self-correction; any hesitation is content-related rather than to find words or grammar - speaks coherently with fully appropriate cohesive features - develops topics fully and appropriately	- uses vocabulary with full flexibility and precision in all topics - uses idiomatic language naturally and accurately	- uses a full range of structures naturally and appropriately - produces consistently accurate structures apart from 'slips' characteristic of native speaker speech	- uses a full range of pronunciation features with precision and subtlety - sustains flexible use of features throughout - is effortless to understand
8	- speaks fluently with only occasional repetition or self-correction; hesitation is usually content-related and only rarely to search for language - develops topics coherently and appropriately	- uses a wide vocabulary resource readily and flexibly to convey precise meaning - uses less common and idiomatic vocabulary skilfully, with occasional inaccuracies - uses paraphrase effectively as required	- uses a wide range of structures flexibly - produces a majority of error-free sentences with only very occasional inappropriacies or basic/non-systematic errors	- uses a wide range of pronunciation features - sustains flexible use of features, with only occasional lapses - is easy to understand throughout; L1 accent has minimal effect on intelligibility
7	- speaks at length without noticeable effort or loss of coherence - may demonstrate language-related hesitation at times, or some repetition and/or self-correction - uses a range of connectives and discourse markers with some flexibility	- uses vocabulary resource flexibly to discuss a variety of topics - uses some less common and idiomatic vocabulary and shows some awareness of style and collocation, with some inappropriate choices - uses paraphrase effectively	- uses a range of complex structures with some flexibility - frequently produces error-free sentences, though some grammatical mistakes persist	shows all the positive features of Band 6 and some, but not all, of the positive features of Band 8
6	- is willing to speak at length, though may lose coherence at times due to occasional repetition, self-correction or hesitation - uses a range of connectives and discourse markers but not always appropriately	- has a wide enough vocabulary to discuss topics at length and make meaning clear in spite of inappropriacies - generally paraphrases successfully	- uses a mix of simple and complex structures, but with limited flexibility - may make frequent mistakes with complex structures, though these rarely cause comprehension problems	- uses a range of pronunciation features with mixed control - shows some effective use of features but this is not sustained - can generally be understood throughout, though mispronunciation of individual words or sounds reduces clarity at times
5	- usually maintains flow of speech but uses repetition, self-correction and/or slow speech to keep going - may over-use certain connectives and discourse markers - produces simple speech fluently, but more complex communication causes fluency problems	- manages to talk about familiar and unfamiliar topics but uses vocabulary with limited flexibility - attempts to use paraphrase but with mixed success	- produces basic sentence forms with reasonable accuracy - uses a limited range of more complex structures, but these usually contain errors and may cause some comprehension problems	shows all the positive features of Band 4 and some, but not all, of the positive features of Band 6
4	- cannot respond without noticeable pauses and may speak slowly, with frequent repetition and self-correction - links basic sentences but with repetitious use of simple connectives and some breakdowns in coherence	- is able to talk about familiar topics but can only convey basic meaning on unfamiliar topics and makes frequent errors in word choice - rarely attempts paraphrase	- produces basic sentence forms and some correct simple sentences but subordinate structures are rare - errors are frequent and may lead to misunderstanding	- uses a limited range of pronunciation features - attempts to control features but lapses are frequent - mispronunciations are frequent and cause some difficulty for the listener
3	- speaks with long pauses - has limited ability to link simple sentences - gives only simple responses and is frequently unable to convey basic message	- uses simple vocabulary to convey personal information - has insufficient vocabulary for less familiar topics	- attempts basic sentence forms but with limited success, or relies on apparently memorised utterances - makes numerous errors except in memorised expressions	shows some of the features of Band 2 and some, but not all, of the positive features of Band 4
2	- pauses lengthily before most words - little communication possible	only produces isolated words or memorised utterances	cannot produce basic sentence forms	speech is often unintelligible
1	- no communication possible - no rateable language			
0	does not attend			

Appendix C. Task Equivalence Checklist

TASK TOPICS		TOPIC					CONTENT					LEXIS					GRAMMAR					FUNCTIONS				
Task Type A	Family	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
	Leisure time	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
	Colour	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
	Festival	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
	Keeping in contact	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
	Dancing	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Task Type B	Describe a friend	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
	Describe a river	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
	Describe a decision	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
	Describe someone in your family	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Task Type C	Qualities of friends	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
	Other relationships	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
	Water-based leisure activities	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
	Economic importance of rivers, lakes	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
	Important decisions	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
	Everyday decisions	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
	Family similarities	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
	Genetic research	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
KEY	Topic											1= Very Familiar					3= Not Sure					5=Very Unfamiliar				
	Content of Questions/Prompts											1=Very Concrete					3= Not Sure					5=Very Abstract				
	Lexis and Grammar (in the questions/prompts)											1= Very Easy					3= Not Sure					5= Very Difficult				
	Lexis, Grammar and Functions elicited by the question/prompts											1=Very Familiar					3= Not Sure					5=Very Unfamiliar				

Task Equivalence Checklist based on Weir, O'Sullivan and Horai (2006:129) and Weir and Wu (2006)

Appendix D. Task Statistics

Task Statistics: Task Type A

TOPICS (Task Type A)	Colour	Dancing	Family	Festivals	Keeping in Contact	Laughing	Leisure Time	Travelling to work
Counts								
Words	52	40	45	40	61	50	49	55
Characters	221	188	184	193	264	217	214	254
Paragraph	5	5	5	5	5	5	5	5
Sentences	7	8	5	5	8	8	7	8
Averages								
Sentences/Paragraph	1.4	1.6	1	1	1.6	1.6	1.4	1.6
Words/Sentence	7.6	5	9	8	7.6	6.2	7	6.9
Characters/Word	4	4.5	4	4.7	4.2	4	4.1	4.3

Readability Statistics: Task Type B

TOPICS (Task Type B)	Describe a Friend	Describe a River	Describe Someone in Family	Describe an Interesting Historic place	Describe a Choice
Counts					
Words	30	40	38	27	44
Characters	132	156	160	129	188
Paragraph	6	6	6	6	6
Sentences	5	5	5	5	5
Averages					
Sentences/Paragraph	1	1	1	1	1
Words/Sentence	5.4	7.4	7	4.8	8.2
Characters/Word	4.2	3.6	4	4.6	4.1
Readability					
Passive Sentences	0%	0%	20%	0%	0%
Flesch-Reading Ease	97.9	100	86.1	75	99.4
Flesch-Kincaid Grade Level	0.9	0.6	2.9	3.9	1.4

Task Statistics: Task Type C

TOPICS (Task Type C)	Qualities of Friends	Other Relationship s	Water- based Leisur e	Economic Importanc e	Family Similarities	Genetic Research	Important Choices	Choices Everyda y Life
Counts								
Words	37	46	51	63	44	34	38	51
Characters	178	234	245	282	210	188	192	219
Paragraph	4	4	4	4	4	4	4	4
Sentences	4	5	3	5	4	4	3	4
Averages								
Sentences/Paragraph	1.3	1.7	1	1.7	1.3	1.3	1	1.3
Words/Sentence	8.5	8.8	16	10.8	10.5	8	12	11.8
Characters/Word	4.6	4.8	4.4	4.3	4.5	5.3	4.8	4.1

Appendix E. Speaking Test Form W

This is Language Proficiency Interview (W). The participant identification number is _____.

Good morning/afternoon/evening

PART I (Task Type A): 4-5 minutes

In this first part I'd like to ask you some questions about some general topics.

Topic A.1 Family

Let's talk about your family.

- Do you have a large family or a small family?
- Can you tell me something about them?
- How much time do you manage to spend with members of your family?
- What sorts of things do you like to do together?

(UCLES, 2002:29)

Topic A.2 :Leisure Time

Now let's talk about your leisure time.

- Do you have any hobbies or interests? [What are they?]
- What is there to do in your free time your city?
- How do you usually spend your holidays?
- Is there anywhere you would particularly like to visit? [Why?]

(UCLES, 2005:80)

PART II (Task Type B): 3-4 minutes

'Now in this second part, I'll give you a topic and I'd like you to talk about it for 1-2 minutes. Before you talk you'll have one minute to think about what you're going to say and you can make some notes if you wish.'

'Here's some paper and pencil for making notes and here's your topic'. [hand in the task card , pencil and paper to participant and wait one minute]

'Now remember, you have 1-2 minutes for this, so don't worry if I stop you. I'll tell you when the time is up. Can you start speaking now please?'

Topic B.1 : Describe one of your friends

Describe one of your friends.

You should say:

How you met

How long you have known each other

How you spend time together

And explain why you like this person.

(UCLES, 2006:77)

PART III (Task Type C) : 4-5 minutes

Topic C.1: Qualities of friends

- What do you think are the most important qualities for friends to have?
- Which are more important to people, their family or their friends? Why?
- What do you think causes friendships to break up?

Topic C.2: Other relationships

- What other types of relationship, apart from friends or family, are important in people's lives today?
- Have relationships with neighbours where you live changed in recent years? How?
- How important do you think it is for a person to spend some time alone? [Why/Why not?]

(UCLES, 2006:77)

PARTICIPANT TASK CARD (W)

Here is your topic. You have one minute to think about what you are going to say and you can make some notes if you wish. You have 1-2 minutes to talk about this topic.

Describe one of your friends.

You should say:

How you met

How long you have known each other

How you spend time together

And explain why you like this person.

Appendix F. Speaking Test Form X

This is an experimental Language Proficiency Interview (X). The participant identification number is__.

Good morning/afternoon/evening

PART I (Task Type A): 4-5 minutes

In this first part I'd like to ask you some questions about some general topics.

Topic A.3: Festivals

Let's talk about festivals.

- Tell me about the most important festival in your country.
- What special food and activities are there in this festival?
- What do you enjoy most about it?
- Do you think festivals are important for a country?

(UCLES, 2002:53)

Topic A.4: Colour

Now let's move on to talk about colour.

- What is your favourite colour? [Why?]
- Do you like the same colours now as you did when you were younger? [Why/Why not?]
- What can you learn about a person from the colours they like?
- Do any colours have a special meaning in your culture?

(UCLES, 2006:54)

PART 2 (Task Type B): 3-4 minutes

‘Now in this second part, I'll give you a topic and I'd like you to talk about it for 1-2 minutes. Before you talk you'll have one minute to think about what you're going to say and you can make some notes if you wish.’

‘Here's some paper and pencil for making notes and here's your topic’. [hand in the task card , pencil and paper to participant and wait one minute]

‘Now remember, you have 1-2 minutes for this, so don't worry if I stop you. I'll tell you when the time is up. Can you start speaking now please?’

Topic B.2 : Describe a river, lake or sea

Describe a river, lake or sea which you like.

You should say:

What the river, lake or sea is called

Where it is

What the land near it is like

And explain why you like this river, lake or sea.

(UCLES, 2005:80)

PART III (Task Type C): 4-5 minutes

Ok now let's move on to some more general questions about this topic. Let's consider:

Topic C.3: Water-based leisure activities

- What do people enjoy doing when they visit rivers, lakes or the sea?
- What benefits do you think people get from the activities they enjoy in the water?
- What are the different advantages and disadvantages of going to the sea or to a swimming pool to enjoy yourself?

Topic C.4: The economic importance of rivers, lakes and the sea

- How does water transport, like boats and ships, compare with other kinds of transport?
- How important is it for a town or city to be located near a river or the sea? Why?
- Have there been any changes in the number of jobs available in fishing and water transport industries, do you think? [Why]

(UCLES, 2005:80)

Participant Task Card (X)

Here is your topic. You have one minute to think about what you are going to say and you can make some notes if you wish. You have 1-2 minutes to talk about this topic.

Describe a river, lake or sea which you like.

You should say:

What the river, lake or sea is called.

Where it is.

What the land near it is like.

And explain why you like this river, lake or sea.

Appendix G. Speaking Test Form Y

This is Language Proficiency Interview (Y). The participant identification number is _____.

Good morning/afternoon/evening

PART 1 (Task Type A): 4-5 minutes

In this first part I'd like to ask you some questions about some general topics.

Topic A.1: Family

Let's talk about your family.

- Do you have a large family or a small family?
- Can you tell me something about them?
- How much time do you manage to spend with members of your family?
- What sorts of things do you like to do together?

(UCLES, 2002:29)

Topic A.2 :Leisure Time

Now let's talk about your leisure time.

- Do you have any hobbies or interests? [What are they?]
- What is there to do in your free time your city?
- How do you usually spend your holidays?
- Is there anywhere you would particularly like to visit? [Why?]

(UCLES, 2005:80)

PART 2 (Task Type B): 3-4 minutes

‘Now in this second part, I'll give you a topic and I'd like you to talk about it for 1-2 minutes. Before you talk you'll have one minute to think about what you're going to say and you can make some notes if you wish.’

‘Here's some paper and pencil for making notes and here's your topic’. [hand in the task card , pencil and paper to participant and wait one minute]

‘Now remember, you have 1-2 minutes for this, so don't worry if I stop you. I'll tell you when the time is up. Can you start speaking now please?’

Task B.3 Describe an important choice you had to make

Describe an important choice you had to make in your life.

You should say:

When you had to make this choice

What you had to choose between

Whether you made a good choice

And explain how you felt when you were making this choice.

(UCLES, 2008: 100)

PART 3 (Task Type C): 4-5 minutes

Ok now let's move on to some more general questions about this topic. Let's consider:

Topic C.5: Important Choices

- What are the typical choices people make at different stages of their lives?
- Should important choices be made by parents rather than by young adults?
- Why do some people like to discuss choices with other people?

Topic C.6: Choices in everyday life

- What kind of choices do people have to make in their everyday life?
- Why do some people choose to do the same things every day? Are there any disadvantages in this?
- Do you think that people today have more choices to make today than in the past?

(UCLES, 2008: 100)

Participant Task Card (Y)

Here is your topic. You have one minute to think about what you are going to say and you can make some notes if you wish. You have 1-2 minutes to talk about this topic.

Describe an important choice you had to make in your life.

You should say:

When you had to make this choice

What you had to choose between

Whether you made a good choice

And explain how you felt when you were making this choice.

Appendix H. Speaking Test Form Z

This is an experimental Language Proficiency Interview (Z). The participant identification number is ____.

Good morning/afternoon/evening

PART 1 (Task Type A): 4-5 minutes

In this first part I'd like to ask you some questions about some general topics.

Topic A.5: Keeping in contact with people

Let's talk about keeping in contact with people.

- How do you usually contact your friends? [Why?]
- Do you prefer to contact different people in different ways? [Why?]
- Do you find it easy to keep in contact with friends and family? [Why/Why not?]
- In your country, did people in the past keep in contact in the same way as they do today?[Why/Why not?]

(UCLES, 2009: 32)

Topic A.6: Dancing

Now let's move on to talk about dancing.

- Do you enjoy dancing? [Why/ Why not?]
- Has anyone ever taught you to dance? [Why/Why not?]
- Tell me about any traditional dancing in your country.
- Do you think that traditional dancing will be popular in the future? [Why/Why not?]

(UCLES, 2008: 32)

PART 1 (Task Type B): 3-4 minutes

‘Now in this second part, I'll give you a topic and I'd like you to talk about it for 1-2 minutes. Before you talk you'll have one minute to think about what you're going to say and you can make some notes if you wish.’

‘Here's some paper and pencil for making notes and here's your topic’. [hand in the task card , pencil and paper to participant and wait one minute]

‘Now remember, you have 1-2 minutes for this, so don't worry if I stop you. I'll tell you when the time is up. Can you start speaking now please?’

Topic B.4 : Describe someone in your family who you like.

Describe someone in your family who you like.

You should say:

How this person is related to you

What this person looks like

What kind of a person he/she is

And explain why you like this person.

(UCLES, 2008: 32)

PART 3 (Task Type C): 4-5 minutes

Ok now let's move on to some more general questions about this topic. Let's consider:

Topic C.7: Family Similarities

- In what ways can people in a family be similar?
- Do you think that daughters are always more similar to mothers than to male relatives? What about sons and fathers?
- In terms of personality, are people more influenced by their family or their friends? In what ways?

Topic C.8: Genetic Research

- Where can people in your country get info about genetic research?
- How do people in your country feel about genetic research?
- Should this research be funded by governments or private companies? Why?

(UCLES, 2008: 32)

Participant Task Card (Z)

Here is your topic. You have one minute to think about what you are going to say and you can make some notes if you wish. You have 1-2 minutes to talk about this topic.

Describe someone in your family who you like.

You should say:

How this person is related to you

What this person looks like

What kind of a person he/she is

And explain why you like this person.

Appendix I. C-tests

Before taking the test, please fill in your details and read the instructions below:

Name:	Nationality:
Gender:	First Language:
If English is not your first language, what do you think your English level is?(please underline)	
Basic Lower Intermediate Intermediate Upper Intermediate Advanced	

Instructions:

The texts in the test have been developed by removing the second (larger) half of every second word in a text. Please reconstruct the original text by filling in the blanks.

Example:

Pablo Picasso was one of the greatest artists of the 20th century. He experi_____ in
ma_____ different sty_____ and cha_____ the wo_____ of a _____ during h_____ time.

Pablo Picasso was one of the greatest artists of the 20th century. He **experimented** in **many**
different **styles** and **changed** the **world** of **art** during **his** time. (English-Online, n.d.)

You have 30 minutes now. Please remember that this is a very difficult test and that you are not expected to fill in all the blanks!

C-test Version I

Text: The Nobel Prize (Text 1)

Nobel Prizes are awards that are given each year for special things that people or groups of people have achieved. They are ¹ awarded in ² six areas ³: physics, chemistry ⁴, medicine, literature ⁵, peace and ⁶ economics. The ⁷ prizes come ⁸ from a fund ⁹ that was ¹⁰ created by ¹¹ the Swedish inventor ¹² Alfred Nobel. He ¹³ wanted to ¹⁴ use so ¹⁵ of his ¹⁶ money to ¹⁷ help make ¹⁸ the world ¹⁹ a better ²⁰ place to ²¹ live. Many ²² organizations determine ²³ who receive ²⁴ the prizes. Prizes can ²⁵ be given ²⁶ to ²⁷ individuals or ²⁸ all round ²⁹, countries and regions ³⁰. Each ³¹ award consists ³² of a gold ³³ medal, a diploma and a lot of money.

Score: _____ / 33 items (Adapted from English-Online, n.d.)

Text: Lack of sleep (Text 2)

For many people, lack of sleep is rarely a matter of choice. Some have ¹ problems getting ² to sleep, others ³ with staying ⁴ asleep until ⁵ the morning ⁶. Despite popular ⁷ belief that ⁸ sleep is ⁹ one of the ¹⁰ most important ¹¹ events, research ¹² shows that ¹³ in an ¹⁴ average night ¹⁵, there are ¹⁶ five stages ¹⁷ of sleep ¹⁸ and four cycles ¹⁹, during which ²⁰ the sequence ²¹ of stages ²² is repeated.

Score: _____ / 22 items (Adapted from UCLES, 2006:109)

Text: Language in science (Text 3)

In Europe, modern science emerged at the same time as the nation state. At first ¹ the scientific ² language of ³ choice remained ⁴ Latin. It was ⁵ scientists that ⁶ communicate with ⁷ other societies ⁸ privileged things ⁹ while protecting ¹⁰ their work ¹¹ from unwelcome ¹² exploitation. Some ¹³, the desire ¹⁴ to protect identity ¹⁵ seemed ¹⁶ to have ¹⁷ been stronger ¹⁸ than the ¹⁹ desire to ²⁰ communicate themselves ²¹, particularly in ²² the case ²³ of mathematicians and doctors ²⁴. In Britain, moreover ²⁵, scientists worried that English had neither the technological vocabulary nor the grammatical resources to express their ideas.

Score: _____ / 25 items

(Adapted from UCLES, 2006: 50)

Text: Student life at Canterbury College (Text 4)

Most of the courses at Canterbury College only take up four days of the week, leaving one day free for independent study. The atmos_____¹ at the coll_____² is th_____³ of a_____⁴ adult envir_____⁵ where a relati_____⁶ of mut_____⁷ respect i_____⁸ encouraged bet_____⁹ students and tut_____¹⁰. Canterbury is a student ci_____¹¹ with sev_____¹² institutes o_____¹³ Higher Education. The city cen_____¹⁴ is ju_____¹⁵ a five min_____¹⁶ walk fr_____¹⁷ the College, eas_____¹⁸ accessible during lu_____¹⁹ or st_____²⁰ breaks. Canterbury College h_____²¹ developed str_____²² international li_____²³ over the ye_____²⁴ and a_____²⁵ a result, many students have the opportunity of visiting and working in a European country in the course of their studies. **Score:** _____ / **25 items** (Adapted from UCLES, 2005: 107)

Text: Taking a gap year (Text 5)

It is quite common these days for young people in many countries to have a break from studying after graduating from high school. The tr_____¹ is n_____² restricted t_____³ rich stud_____⁴ who ha_____⁵ the mo_____⁶ to tra_____⁷, but i_____⁸ also e_____⁹ among poo_____¹⁰ students w_____¹¹ choose t_____¹² work and bec_____¹³ economically indep_____¹⁴ for a per_____¹⁵ of ti_____¹⁶. The rea_____¹⁷ for th_____¹⁸ trend m_____¹⁹ involve the recog_____²⁰ that a yo_____²¹ adult w_____²² passes dire_____²³ from sch_____²⁴ to unive_____²⁵ is rat_____²⁶ restricted i_____²⁷ terms o_____²⁸ general know_____²⁹ and exper_____³⁰ of the wo_____³¹. By con_____³², those who have spent some time earning a living or traveling to other places have a broader view of life and better personal resources to draw on. **Score:** _____ / **32 items** (Adapted from UCLES, 2006: 165)

C-test Version II

Text: The Nobel Prize (Text 1)

Nobel Prizes are awards that are given each year for special things that people or groups of people have achieved. They a_____¹ awarded i_____² six ar_____³: physics, chem_____⁴, medicine, liter_____⁵, peace a_____⁶ economics. T_____⁷ prizes co_____⁸ from a fu_____⁹ that w_____¹⁰ created b_____¹¹ the Swedish inve_____¹² Alfred Nobel. H_____¹³ wanted t_____¹⁴ use so_____¹⁵ of h_____¹⁶ money t_____¹⁷ help ma_____¹⁸ the wo_____¹⁹ a bet_____²⁰ place t_____²¹ live. Ma_____²² organizations dete_____²³ who rece_____²⁴ the prizes. Prizes c_____²⁵ be gi_____²⁶ t_____²⁷ individuals o_____²⁸ all ra_____²⁹, countries and reli_____³⁰. Ea_____³¹ award cons_____³² of a go_____³³ medal, a diploma and a lot of money. **Score:** _____ / **33 items** (Adapted from English-Online, n.d.)

Text: Street Art (Text 6)

Street Art is a very popular form of art that is spreading quickly all over the world. You c¹ find i² on buil³, sidewalks, str⁴ signs a⁵ trash ca⁶ from Tokyo t⁷ Paris. Street art has bec⁸ a global cul⁹ and ev¹⁰ art muse¹¹ and gall¹² are colle¹³ the wo¹⁴ of street art¹⁵. Street art started o¹⁶ very secr¹⁷ because i¹⁸ is ill¹⁹ to pa²⁰ on public a²¹ private prop²² without permi²³. People of²⁴ have diff²⁵ opinions ab²⁶ street art. So²⁷ think i²⁸ is a cr²⁹ and oth³⁰ think i³¹ is a ve³² beautiful, n³³ form of culture. **Score:** ____ / 33 (Adapted from English-Online, n.d.)

Text: Minority languages (Text 7)

There are currently approximately 6,800 languages in the world. This gr¹ variety o² languages ca³ about lar⁴ as a res⁵ of geogra⁶ isolation. B⁷ in tod⁸ world, fac⁹ such a¹⁰ government initi¹¹ and econ¹² globalization a¹³ contributing t¹⁴ a huge decr¹⁵ in t¹⁶ number of languages. O¹⁷ factor wh¹⁸ may he¹⁹ to ens²⁰ that so²¹ endan²² languages d²³ not d²⁴ out compl²⁵ is peo²⁶ increasing apprec²⁷ of th²⁸ cultural iden²⁹. This has been encouraged through programs of language classes for children and through 'apprentice' schemes. **Score:** ____ / 29 (Adapted from UCLES, 2004:44)

Text: Taking a gap year (Text 5)

It is quite common these days for young people in many countries to have a break from studying after graduating from high school. The tr¹ is n² restricted t³ rich stud⁴ who ha⁵ the mo⁶ to tra⁷, but i⁸ also evi⁹ among poo¹⁰ students w¹¹ choose t¹² work and bec¹³ economically indep¹⁴ for a per¹⁵ of ti¹⁶. The rea¹⁷ for th¹⁸ trend m¹⁹ involve the recog²⁰ that a yo²¹ adult w²² passes dire²³ from sch²⁴ to unive²⁵ is rat²⁶ restricted i²⁷ terms o²⁸ general know²⁹ and exper³⁰ of the wo³¹. By con³², those who have spent some time earning a living or traveling to other places have a broader view of life and better personal resources to draw on. **Score:** ____ / 32 items (Adapted from UCLES, 2006: 165)

Text: Children's books (Text 8)

Everyone has a favourite children's book. Fond memo¹ remain o² books th³ we ha⁴ read a⁵ re-read, transp⁶ the rea⁷ back t⁸ childhood. N⁹ matter i¹⁰ they've lo¹¹ the ha¹² of rea¹³ in la¹⁴ life; i¹⁵ is a ra¹⁶ adult w¹⁷ does n¹⁸ have a resi¹⁹ tenderness f²⁰ a defining wo²¹, whether it's '*the Story of Babar, A Bear Called Paddington* or *Ballet Shoes*'. Some²² it's the te²³ that attr²⁴ but mo²⁵ often th²⁶ not, it's the illust²⁷ that dr²⁸ in the young reader. **Score:** ____ / 28 items (Adapted from Time-Out, 2007)

C-test Version III

Text: The Nobel Prize (Text 1)

Nobel Prizes are awards that are given each year for special things that people or groups of people have achieved. They are ¹ awarded to ² six areas ³: physics, chemistry ⁴, medicine, literature ⁵, peace and ⁶ economics. These ⁷ prizes come ⁸ from a fund ⁹ that was ¹⁰ created by ¹¹ the Swedish inventor ¹² Alfred Nobel. He ¹³ wanted to ¹⁴ use so ¹⁵ of his ¹⁶ money to ¹⁷ help make ¹⁸ the world ¹⁹ a better ²⁰ place to ²¹ live. Many ²² organizations determine ²³ who receive ²⁴ the prizes. Prizes can ²⁵ be given ²⁶ to ²⁷ individuals or ²⁸ all round ²⁹, countries and religions ³⁰. Each ³¹ award consists ³² of a gold ³³ medal, a diploma and a lot of money.

Score: _____ / 33 items (Adapted from English-Online, n.d.)

Text: Street Art (Text 6)

Street Art is a very popular form of art that is spreading quickly all over the world. You can ¹ find it ² on buildings ³, sidewalks, street ⁴ signs and ⁵ trash cans ⁶ from Tokyo to ⁷ Paris. Street art has become ⁸ a global cultural ⁹ and even ¹⁰ art museums ¹¹ and galleries ¹² are collecting ¹³ the works ¹⁴ of street art ¹⁵. Street art started out ¹⁶ very secret ¹⁷ because it ¹⁸ is illegal ¹⁹ to paint ²⁰ on public and ²¹ private property ²² without permission ²³. People of ²⁴ have different ²⁵ opinions about ²⁶ street art. So ²⁷ think it ²⁸ is a crime ²⁹ and others ³⁰ think it ³¹ is a very ³² beautiful new ³³ form of culture. **Score:** _____ / 33 (Adapted from English-Online, n.d.)

Text: Noise (Text 9)

In general, it is plausible to suppose that we should prefer peace and quiet to noise. And you ¹ most of ² us have ³ had the ⁴ experience of ⁵ having to ⁶ adjust to ⁷ sleeping in ⁸ the mountains ⁹ or the ¹⁰ countryside because ¹¹ it was ¹² initially 'too ¹³ quiet': a ¹⁴ experience that ¹⁵ suggests that ¹⁶ humans are ¹⁷ capable of ¹⁸ adapting to ¹⁹ a wide range ²⁰ of noise ²¹ levels. Research supports ²² this view ²³. **Score:** _____ / (23 items) (Adapted from UCLES, 2009: 96)

Text: Language in science (Text 3)

In Europe, modern science emerged at the same time as the nation state. At first ¹ the scientific ² language of ³ choice remained ⁴ Latin. It took ⁵ scientists to ⁶ communicate with ⁷ other societies ⁸ privileged thinkers ⁹ while protecting ¹⁰ their works ¹¹ from unwelcome ¹² exploitation. Some ¹³, the desire ¹⁴ to protect identity ¹⁵ seemed ¹⁶ to have ¹⁷ been stronger ¹⁸ than the ¹⁹ desire to ²⁰ communicate with ²¹ particularly in ²² the case ²³ of mathematicians and doctors ²⁴. In Britain, more ²⁵, scientists worried that English had neither the technological vocabulary nor the grammatical resources to express their ideas. **Score:** _____ / 25 items (Adapted from UCLES, 2006: 50)

Text: History of early cinema (Text 10)

The history of the cinema in its first thirty years is one of major and, to this day, unparalleled expansion of growth. Beginning a _____¹ something unu _____² in a han _____³ of b _____⁴ cities – New York, London, Paris and Berlin, the n _____⁵ medium qui _____⁶ found i _____⁷ way acr _____⁸ the wo _____⁹, attracting lar _____¹⁰ and lar _____¹¹ audiences wher _____¹² it w _____¹³ shown and repl _____¹⁴ other fo _____¹⁵ of entert _____¹⁶ as i _____¹⁷ d _____¹⁸ so. A _____¹⁹ audiences gr _____²⁰, so d _____²¹ the pla _____²² where fi _____²³ were sh _____²⁴. Meanwhile, films thems _____²⁵ developed fr _____²⁶ being sh _____²⁷ attractions on _____²⁸ a couple of minutes lo _____²⁹, to the full-length feature that has dominated the world's screens up to the present time. **Score: _____ / 29 items** (Adapted from UCLES, 2005: 111)

Appendix J. Invitation (Language Schools)

Dear Sir/Madam,

My name is Nahal Khabbazzbashi and I am a 2nd year PhD. student at Oxford University. I would hereby like to invite you for collaboration in a research project that I am conducting. The focus of the project is on the speaking performance of Iranian second language learners of English in assessment contexts and particularly in standardised tests such as IELTS.

The project would involve the administration of a short general proficiency test and two IELTS speaking tests (each lasting between 10 and 15 minutes) followed by short questionnaires to a sample of students at your institute. The study is targeted for adult Iranian students of English who are preparing to take the IELTS exam in the near future. The research will provide an excellent opportunity for the selected students to practise their speaking skills with a trained examiner (I am a certified IELTS examiner). This study has also received ethical approval from Oxford's Central University Research Ethics Committee (see attached).

The data collection is planned for end of July/beginning of August 2010 (the first two weeks of Mordad). If your esteemed organization is interested in participating in the study, I would appreciate if you could inform me by return e-mail as soon as possible so that further arrangements could be made. Your collaboration will be highly appreciated.

In case of any questions please do not hesitate to contact me. I thank you in advance for your kind consideration and look forward to hearing back from you.

Kind Regards,

Nahal Khabbazzbashi

Appendix K. Participant Information Sheet

Practice Tests for the IELTS Speaking Module: Share Your Experience

Information for Participants

Invitation

You are being invited to take part in a research study. Before you decide to participate, it is important to understand why the research is being conducted and what your participation entails. Please take time to read the following information carefully. Please ask if there are any aspects of the project that are unclear or if you would like more information. Take time to decide whether or not you would like to take part in this research.

Research project and purpose

Let me introduce myself. My name is Nahal Khabbazzbashi and I am a doctoral student reading for a DPhil (PhD) in Education at the University of Oxford. My general area of interest is the field of English language testing with specific focus on the IELTS (International English Language Testing System) speaking section. The results of my research will be the basis for my DPhil thesis and will be under the co-supervision of Dr. Robert Vanderplank and Dr. Catherine Walter. The provisional title for my research project is: 'But I had nothing to say: building a topic validity argument for the IELTS Speaking Test'.

In this project would like to explore your personal experience with the IELTS speaking test and will ask you to share your thoughts, comments and feedback with me. For this purpose the project will consist of the following:

- A short English grammar test
- Two practice IELTS speaking tests (from already published IELTS material; each lasting about 11-14 minutes to be either audio/video-recorded)
- Questionnaires about your English language background, your experience with the two tests and a comparison of the two tests
- A brief interview on your thoughts about the tests

The above process will take a maximum of two hours (which will take place at a time of your convenience) and in return for you time and participation I will provide you with feedback about your speaking performance and some (hopefully) useful tips for taking the IELTS. You should note however that this study is an independent research project and will have no effects whatsoever on your scores or results in the authentic IELTS tests. Please also know all data will be anonymous and your name will not be appearing anywhere in the project but will be referred to by an identification number only. No-one will have access to your data except for me (and raters who will rate your anonymous speaking tests) and nothing will be traced back to you. You have the right to withdraw from the research anytime you wish to by simply informing me.

Why have you been chosen?

This study is focusing on foreign language learners of English who are interested in taking the IELTS or have recently taken the test. You will therefore be probably familiar with the general format of the test and what it concerns. If you have taken IELTS preparation classes before you can identify the materials, books and tips you found useful and how relevant they are/were in taking the test. You can inform me of your personal experience of this test and your contribution in sharing your experience will assist me in researching this particular area of standardized language testing.

Do I have to take part? Are there any benefits?

It is your decision to take part in this study. You can decide to stop participating at any time. You do not need to answer questions that you do not wish to. Every effort will be made to preserve confidentiality. The benefits are your contribution to studying standardized English language tests. This study also allows for additional practice and to further familiarize yourself with the format of the IELTS speaking test and what it involves and will therefore be useful in your preparations. As the speaking tests will be practice tests only, not taking part in the study will not disadvantage you in anyway as you can refer to already published material (which will also be used in this study) for further practice/information.

What will happen to the results of this research?

The results of this research will form the basis of my PhD thesis. The study will take place over the next 24 months after which any published results will be publicly available however any data included is anonymous. The original collected data will be securely locked away and no one will have further access to them (except for me). Once all work on the thesis is completely finished, the data will be destroyed.

Who is funding and organizing this research?

This research is currently self-funded and is organized as an independent DPhil thesis research project in conjunction with the Department of Education, Oxford University. The project has also been reviewed by, and received ethics clearance through, the University of Oxford Central University Research Ethics Committee.

Contact for Further Information or Follow-up

Should you have any further questions about this research, please feel free to contact: Nahal Khabbazzbashi, Department of Education, 15 Norham Gardens, Oxford, UK OX2 6PY. Via e-mail at nahal.khabbazzbashi@education.ox.ac.uk. My supervisor, Dr. Robert Vanderplank has overall control of the project and every step of the research will be subject to his approval. His mailing address is Dr Robert Vanderplank, Director, Oxford University Language Centre, 12 Woodstock Road, Oxford OX2 6HT, United Kingdom and his e-mail is robert.vanderplank@lang.ox.ac.uk. My co-supervisor is Dr. Catherine Walter whose mailing address is Dr Catherine Walter, Department of Education, 15 Norham Gardens, Oxford OX2 6PY, United Kingdom; e-mail: Catherine.walter@education.ox.ac.uk. Should you have any further information, insight into these issues, or questions your inquiries are most welcome.

Thank you for taking the time to read this information sheet.

Appendix L. Participant Consent Form

‘But I had nothing to say’-Building a topic validity argument for the IELTS Speaking Test

Speaking test/Interview/questionnaire consent form

The research project seeks to explore the effects of conducting two different IELTS practice speaking tests and participants’ individual experiences. This is a study undertaken by Nahal Khabbazzashi, a doctoral student at the Department of Education, University of Oxford and involves a grammar test, two speaking tests, questionnaires and a brief interview.

1. I have read and understood the information about this study and have had the opportunity to ask questions.
2. I understand that I can withdraw from the study without consequence at anytime simply by informing the researcher of my decision.
3. I understand who will have access to identifying information provided and what will happen to the data at the end of the project.
4. I am aware of who to contact should I have questions following my participation in this study.
5. I understand that this project has been reviewed by and received ethical clearance through the University of Oxford Central University Research Ethics Committee.

I agree to participate in this study.

Name: _____

Date: _____

Signature: _____

Researcher: _____

Date: _____

Signature: _____

Appendix M. Invitation (Raters)

Dear Colleague,

My name is Nahal and I am a 3rd year DPhil student at the Department of Education, University of Oxford. The purpose of this e-mail is to invite you for collaboration in my research project.

The focus of my research is on the effects of topic on the speaking performance of non-native speakers of English in language proficiency interviews. The format of the language proficiency interview that I have used for collecting my data is that of the IELTS speaking test. I have administered two speaking tests to my participants (each test consisting of five different topics/tasks) and now these tasks need to be rated and scored. This is where I need your help.

The rating process would include the following steps:

- Filling a short 'task equivalence' checklist
- Listening to a number of different tasks, giving each task a score (using a rating scale)
- Tallying the different functions that you hear in each task (using an observation checklist)
- Commenting (if necessary) on tasks that you found particularly difficult to rate or aspects of the rating process that you found particularly problematic.
- Filling a short 'task equivalence' checklist

It is preferable if you have some familiarity with the IELTS speaking test. Can you also please tell me if you have previous teaching/examining experience?

According to my calculations, each rater would need a total of approximately 25-28 hours for the whole process. I have received a small grant which allows me to pay a flat rate of £200 per rater for completing the job. You will have between 10 to 14 days to finish the ratings. Ideally, I would like to have the ratings completed before the beginning of Trinity Term but I also understand that you may have assignments/exams in which case we can talk about the dates that are most convenient for you but that also fit in with my schedule. Please therefore let me know of your availability.

Before starting the ratings, you will receive an IELTS DVD which explains the rating process with more detailed examples. I will also provide you with a rating scale and the checklists so that you can familiarise yourself with the instruments. We would then have a meeting to discuss the rating process and any problems/issues before the main ratings start.

It would be great if you could get back to me as soon as possible.

Kind Regards,

Nahal Khabbazzbashi

Appendix N. Topic Development Scale (ETS, 2004)

<i>Topic Development</i>	<i>Score</i>
Develops topic fully and appropriately Response is sustained and relevant to the task It is generally well developed and coherent	5
Develops topic coherently with a fair amount of elaboration and specificity Ideas are well organized and relevant to the task	4
Response is mostly coherent and sustained and conveys relevant ideas/information Overall development is somewhat limited and usually lacks elaboration and specificity Relationships between ideas may at times not be immediately clear	3
The response is connected to the task, though the number of ideas presented or the development of ideas is limited.. Mostly basic ideas are expressed with limited elaboration (details and support) Connections of ideas may be unclear	2
Limited relevant content is expressed. The response generally lacks substance beyond expression of very basic ideas. At times relevant substance may be vaguely expressed or repetitious. Speaker may be unable to sustain speech to complete task and may rely heavily on repetition of the prompt.	1

Based on the TOEFL iBT Independent Speaking Rubrics: Topic Development Scale (ETS, 2004)

Appendix O. Observation Checklist (Additional Details)

Observation Checklist: Details	
<i>Informational Functions</i>	
Providing personal information	Give information on present circumstances, past experiences or future plans
Justifying Opinions:	Express reasons for assertions he/she had made
Suggesting:	Suggest a particular idea
Comparing:	Compare things/people/events
Contrasting:	Contrast things/people/events
Describing:	Describe a sequence of events Describe a scene
Elaborating:	Elaborate on, or modify an opinion
Summarising:	Summarize what he/she has said
Staging:	Separate out or interpret the parts of an issue
<i>Interactional functions</i>	
Negotiating meaning	Check understanding Ask for clarification when an utterance is misheard or misinterpreted
Agreeing	Agree with a question/assertion (apart from 'yes', 'yeah' or nonverbal)
Disagreeing	Disagree with a question (apart from 'no' or nonverbal)
Comment on task/topic difficulty or unfamiliarity	e.g. I don't know anything about this topic e.g. That's a difficult question
Asking for task/topic clarification	e.g. Should I talk about 'x' or 'y'?

Adapted from O'Sullivan, Weir and Saville (2002:54)

Appendix P. Sample Score Sheet

Task/Speaker Code		
Section 1 (Speaking Band Descriptors Scores)		
Fluency and Coherence		
Lexical Resource		
Grammatical Accuracy and Range		
Pronunciation		
Section 2 (Topic Development Scale Score)		
Topic Development		
Section 3 (Tally of Functions)		
Providing Personal Information		
Providing Non-Personal Information		
Expressing Opinions		
Justifying Opinions		
Explaining		
Suggesting		
Expressing a Preference		
Comparing		
Contrasting		
Paraphrasing		
Narrating		
Describing		
Elaborating		
Summarizing		
Paraphrasing		
Speculating		
Staging		
Analyzing		
Negotiation of Meaning		
Conversation Repair		
Self-Repair		
Agreeing		
Disagreeing		
Comment on task/topic difficulty		
Ask for task/topic clarification		
Any functions not included above (give examples in 'comments')		
Section 4 (Comments)		

Observation 'Functions' Checklist (Adapted from Weir, O'Sullivan and Saville (2002))

Appendix Q. Score Sheets: Instruction to Raters

General

- For each test item, first write down the item/speaker reference number, which is the same as the audio file number.
- In order to avoid remarking the same item, make sure to have two separate folders for 'to be rated' and 'completed rating' files and as soon as rating an item is finished transfer the audio file from 'to be rated' to 'completed rating' folder.
- The reference number is 10 digits so it is easier to copy/paste the number to your excel worksheet rather than typing it in.
- The score sheet is divided into three sections. For each section, you need to refer to the relevant document (see below). There is also an additional 'comments' section in the very end.

Scoring Spoken Test Items

- Refer to the 'Rating Scale' document and on the basis of the descriptors, assign a band score across the four criteria of 'Fluency and Coherence, Lexical Resource, Grammatical Range and Accuracy and Pronunciation' for every performance item.
- You can pause the audio file at any time and you can listen to each task item more than once if necessary
- If you are unsure of a rating, jot down the score that you think best captures the performance but highlight the cell in yellow. This way the performance can be flagged for double marking. A separate section is also provided for 'comments' so if you come across any problems or issues or if you just want to comment on the rating process please feel free to use that section. It will be very welcome!
- Some audio files may include noise (e.g. telephone ringing, doors opening, construction work in the background). These were beyond my control. The speech should still be intelligible but if you encounter audio files in which you have a lot of trouble hearing the test taker, highlight the item reference number in red.

Topic Development

- Refer to the ‘Topic Development’ scale and on the basis of the descriptors, give each task item a score between 1 and 5.

Checklist of Functions

- Read the Weir, O’Sullivan and Saville (2002) article provided in your pack. This will help you familiarise yourself with the purpose of the checklist.
- Refer to the ‘Observation Checklist’ document for more details/examples of the functions.
- Each time you hear a *different type* of function, mark it down. You can pause the audio at any time to mark down your observations.
- If you observe particular functions that are not included in the list, take a note of them with examples in the ‘comments’ section.

Comments

- There is an additional ‘comments’ section in which you can include comments regarding any aspect of the rating process. In order to help save time you can use the following abbreviations if you are referring to any of the categories below:
 - **FC:** Fluency and Coherence
 - **LR:** Lexical Resource
 - **GRA:** Grammatical Range and Accuracy
 - **P:** Pronunciation
 - **TD:** Topic Development
 - **OC:** Observation Checklist
 - **F:** Function(s)

Appendix R. Examinee Measurement Report

ID	OBSVD AVG	FAIR-M AVG	MEAS URE	MODEL S.E	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
65	3.9	3.92	-3.37	0.12	0.91	-0.8	0.89	-0.9
13	4.1	4.03	-3.1	0.15	0.99	0	0.98	0
66	4.2	4.05	-3.06	0.2	1.13	0.7	1.21	1
63	4	4.08	-2.99	0.21	0.57	-2.5	0.59	-2.2
82	4.3	4.2	-2.74	0.15	0.74	-1.8	0.74	-1.7
78	4.3	4.24	-2.66	0.15	0.85	-0.9	0.84	-1
68	4.4	4.42	-2.33	0.17	0.82	-1.1	0.81	-1.1
5	4.4	4.43	-2.31	0.19	1.55	2.5	1.44	2
60	4.5	4.44	-2.3	0.14	0.83	-1.2	0.79	-1.5
71	4.4	4.45	-2.27	0.1	0.76	-2.5	0.75	-2.5
61	4.5	4.5	-2.18	0.15	1.3	1.9	1.34	2
49	4.8	4.58	-2.05	0.18	0.46	-3.6	0.46	-3.5
80	4.6	4.75	-1.77	0.16	1.03	0.2	1.06	0.4
64	4.9	4.84	-1.63	0.14	0.75	-1.8	0.74	-1.8
19	4.9	4.87	-1.57	0.11	0.8	-1.8	0.8	-1.8
3	4.9	4.9	-1.52	0.14	0.83	-1.1	0.84	-1.1
24	4.9	4.91	-1.51	0.1	0.81	-1.8	0.81	-1.8
4	4.8	4.93	-1.47	0.19	0.92	-0.3	0.91	-0.4
30	4.8	4.96	-1.43	0.16	1.24	1.4	1.3	1.7
38	4.9	4.96	-1.43	0.18	1.18	0.9	1.21	1
77	5	4.97	-1.42	0.13	0.8	-1.5	0.81	-1.4
34	4.9	4.97	-1.42	0.1	1.21	1.9	1.22	2
79	5	4.99	-1.37	0.13	0.82	-1.3	0.83	-1.2
6	4.9	5	-1.35	0.18	1.2	1	1.2	1
28	5.1	5.02	-1.34	0.14	0.79	-1.4	0.8	-1.4
67	4.9	5.04	-1.3	0.18	0.89	-0.5	0.89	-0.5
35	5.1	5.1	-1.2	0.1	1	0	1	0
41	5.1	5.12	-1.17	0.1	0.85	-1.5	0.85	-1.5
55	5.2	5.15	-1.12	0.14	0.7	-2.2	0.7	-2.1
7	5.4	5.15	-1.11	0.18	1	0	0.98	0
81	5.4	5.16	-1.11	0.18	0.8	-1	0.79	-1.1
36	5.3	5.16	-1.1	0.14	0.82	-1.2	0.82	-1.2
75	5.2	5.17	-1.09	0.18	0.96	-0.1	0.96	-0.1
52	5.1	5.22	-1	0.18	1.2	1	1.2	1
31	5.4	5.23	-0.98	0.18	0.64	-2.1	0.64	-2.1
74	5.3	5.3	-0.87	0.12	0.75	-2.1	0.74	-2.1
44	5.3	5.35	-0.8	0.1	0.98	-0.2	0.97	-0.2
47	5.4	5.35	-0.79	0.16	0.72	-1.8	0.71	-1.9
14	5.6	5.36	-0.78	0.17	1.23	1.2	1.22	1.2
9	5.5	5.38	-0.74	0.16	1.03	0.2	1.01	0.1
42	5.4	5.4	-0.71	0.1	0.85	-1.5	0.84	-1.6
37	5.4	5.41	-0.69	0.16	0.89	-0.6	0.88	-0.7

ID	OBSVD AVG	FAIR-M AVG	MEAS URE	MODEL S.E	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
57	5.2	5.41	-0.69	0.18	1.15	0.8	1.14	0.7
29	5.6	5.42	-0.68	0.16	0.85	-0.8	0.83	-1
62	5.4	5.46	-0.61	0.16	1.19	1.1	1.18	1
27	5.5	5.46	-0.6	0.12	0.78	-1.9	0.77	-2
51	5.4	5.48	-0.56	0.16	0.85	-0.8	0.85	-0.9
69	5.6	5.5	-0.53	0.12	0.7	-2.6	0.68	-2.8
73	5.5	5.52	-0.5	0.14	1.03	0.2	1.07	0.4
45	5.6	5.54	-0.47	0.15	0.73	-1.9	0.74	-1.8
50	5.6	5.57	-0.42	0.13	1.49	3.2	1.5	3.1
48	5.6	5.61	-0.33	0.18	0.66	-1.9	0.64	-2
21	5.6	5.61	-0.33	0.12	0.91	-0.6	0.9	-0.8
12	5.5	5.63	-0.3	0.18	1.31	1.5	1.31	1.5
43	5.8	5.71	-0.17	0.16	1.08	0.5	1.05	0.3
11	5.5	5.72	-0.14	0.18	1.15	0.8	1.16	0.8
20	5.7	5.74	-0.1	0.19	0.9	-0.4	0.88	-0.5
10	6	5.76	-0.07	0.19	0.82	-0.8	0.77	-1.1
53	5.8	5.76	-0.06	0.15	0.89	-0.6	0.87	-0.8
18	5.8	5.78	-0.03	0.15	0.9	-0.5	0.89	-0.6
2	5.7	5.8	0	0.19	0.61	-2.2	0.59	-2.3
26	5.9	5.82	0.05	0.12	1.05	0.4	1.04	0.3
76	6	5.84	0.09	0.17	0.73	-1.6	0.69	-1.8
54	5.9	5.86	0.12	0.15	0.9	-0.6	0.89	-0.6
56	5.8	5.89	0.17	0.19	0.89	-0.5	0.87	-0.6
17	6.1	5.91	0.21	0.17	1.07	0.4	1.06	0.3
23	6	5.96	0.3	0.14	1.46	2.7	1.4	2.4
16	6	6	0.38	0.1	1.24	2.2	1.25	2.2
46	6.1	6.02	0.43	0.11	1.02	0.2	1	0
22	6	6.04	0.46	0.11	1.13	1.1	1.17	1.4
59	6.1	6.15	0.66	0.17	1.44	2.2	1.49	2.4
33	6.2	6.2	0.76	0.16	0.9	-0.6	0.94	-0.3
39	6.2	6.22	0.8	0.17	0.69	-1.8	0.72	-1.6
58	6.7	6.6	1.56	0.19	0.97	0	1.07	0.4
15	6.6	6.65	1.65	0.15	1.56	3.2	1.88	4.7
70	6.8	6.7	1.75	0.17	0.96	-0.1	1.28	1.5
32	6.7	6.72	1.79	0.19	1.02	0.1	1.3	1.5
8	6.8	6.85	2.05	0.17	2.98	7.9	3.97	9
1	7.2	7.23	2.91	0.15	2.15	6.2	2.83	7.5
72	7.3	7.33	3.17	0.17	1.3	1.6	2.07	4.1
40	7.4	7.42	3.44	0.2	1.1	0.5	3.56	6.5
Mean (N=81)	5.4	5.41	-0.63	0.15	1.00	-1	1.07	.0
S.D	0.7	0.75	1.36	.03	.35	1.9	.55	2.2
Model, Sample: RMSE .16 Adj (True) S.D. 1.36 Separation 8.64 Strata 11.86 Reliability .99 Model, Fixed (all same) chi-square: 6043.5 d.f.: 80 significance (probability): .00								

Examinee Measurement Report (FullA)

Appendix S. Table of Unexpected Responses

CAT	SCORE	EXP.	RESD	STRES	EXAMINEE	RATER	TOPIC	CRITERION
3	3	5	-2	-9	40	R3	A.1	TD
3	3	5	-2	-9	40	R3	A.2	TD
2	2	4.9	-2.9	-8.8	1	R3	C.6	TD
3	3	4.9	-1.9	-8.5	72	R3	C.6	TD
2	2	4.9	-2.9	-8.4	1	R1	B.1	TD
2	2	4.9	-2.9	-8.1	8	R3	A.6	TD
1	1	4.7	-3.7	-7.3	8	R4	A.6	TD
2	2	4.7	-2.7	-5.7	1	R4	C.6	TD
2	2	4.7	-2.7	-5	8	R2	A.6	TD
3	3	4.8	-1.8	-4.8	8	R3	A.3	TD
2	2	4.6	-2.6	-4.5	15	R1	A.1	TD
3	3	4.8	-1.8	-4.5	70	R3	A.6	TD
4	4	5	-1	-4.5	40	R3	C.7	TD
4	4	5	-1	-4.5	72	R3	B.4	TD
3	3	4.8	-1.8	-4.3	15	R3	A.6	TD
3	3	4.8	-1.8	-4.3	32	R3	A.3	TD
3	3	4.8	-1.8	-4.2	70	R3	C.6	TD
3	3	4.8	-1.8	-4.2	72	R4	C.6	TD
3	3	4.8	-1.8	-4.1	15	R3	A.1	TD
4	4	7.4	-3.4	-3.8	1	R1	B.3	P
2	2	4.5	-2.5	-3.8	15	R4	A.3	TD
6	6	7.9	-1.9	-3.7	1	R1	A.1	LR
3	3	4.8	-1.8	-3.7	58	R3	C.6	TD
3	3	4.7	-1.7	-3.6	32	R3	C.8	TD
6	6	7.9	-1.9	-3.3	1	R1	B.3	GA
3	3	4.7	-1.7	-3.3	8	R1	A.3	TD
4	4	4.9	-0.9	-3.3	40	R4	A.6	TD
2	2	4.3	-2.3	-3.2	17	R3	C.4	TD
1	1	3.5	-2.5	-3.1	29	R3	C.4	TD
4	4	6	-2	-3.1	59	R1	A.1	P
3	3	5.3	-2.3	-3	44	R1	B.1	P
4	4	5.9	-1.9	-3	39	R1	A.3	P
1	1	3.3	-2.3	-2.9	7	R3	B.2	TD
6	6	8.1	-2.1	-2.9	72	R3	B.4	P
4	4	5.9	-1.9	-2.9	16	R1	B.1	P
4	4	5.9	-1.9	-2.9	22	R1	A.1	P
4	4	5.9	-1.9	-2.9	22	R1	A.2	P
4	4	5.9	-1.9	-2.9	33	R1	B.2	P
4	4	5.9	-1.9	-2.9	59	R1	A.4	P
4	4	5.8	-1.8	-2.9	33	R1	C.1	P
1	1	3.2	-2.2	-2.8	81	R3	C.4	TD
4	4	6.2	-2.2	-2.8	73	R2	A.2	LR
2	2	4.1	-2.1	-2.8	46	R3	C.6	TD

Cat	Score	Exp.	Resd	StRe	Examinee	Rater	Topic	Criterion
5	5	7	-2	-2.8	16	R1	A.1	FC
3	3	4.6	-1.6	-2.8	58	R3	A.6	TD
4	4	6.1	-2.1	-2.7	27	R4	B.3	GA
4	4	6.1	-2.1	-2.7	44	R2	B.1	LR
5	5	6.9	-1.9	-2.7	16	R1	A.2	FC
6	6	7.6	-1.6	-2.7	8	R2	A.6	FC
6	6	7.6	-1.6	-2.7	8	R2	A.6	LR
3	3	5	-2	-2.6	36	R4	C.7	P
2	2	4	-2	-2.6	20	R3	C.6	TD
2	2	4	-2	-2.6	46	R1	C.6	TD
4	4	5.7	-1.7	-2.6	16	R1	A.4	P
4	4	5.7	-1.7	-2.6	22	R1	C.1	P
4	4	5.7	-1.7	-2.6	22	R1	B.2	P
4	4	5.7	-1.7	-2.6	26	R1	B.4	P
4	4	5.7	-1.7	-2.6	59	R4	C.3	P
3	3	4.5	-1.5	-2.6	15	R1	C.6	TD
3	3	4.5	-1.5	-2.6	15	R4	A.1	TD
4	4	4.9	-0.9	-2.6	1	R1	A.6	TD
2	2	4	-2	-2.4	69	R3	A.6	TD
4	4	5.7	-1.7	-2.2	17	R1	C.5	P
4	4	5.7	-1.7	-2.1	22	R1	C.3	P
8	8	6	2	2.6	26	R2	C.6	GA
8	8	6	2	2.6	50	R2	A.4	GA
8	8	6	2	2.6	62	R1	C.7	GA
5	5	3	2	2.6	35	R4	B.1	TD
5	5	3	2	2.6	67	R1	A.1	TD
5	5	2.9	2.1	2.6	37	R2	C.1	TD
5	5	2.9	2.1	2.6	67	R1	A.3	TD
8	8	5.9	2.1	2.7	6	R1	A.2	LR
8	8	5.9	2.1	2.7	34	R3	B.2	GA
5	5	2.9	2.1	2.7	3	R1	B.1	TD
5	5	2.9	2.1	2.7	6	R4	B.1	TD
5	5	2.9	2.1	2.7	6	R4	B.3	TD
5	5	2.9	2.1	2.7	41	R4	A.3	TD
5	5	2.9	2.1	2.7	44	R2	C.1	TD
5	5	2.9	2.1	2.7	52	R4	B.2	TD
9	9	6.7	2.3	2.7	8	R4	A.6	P
6	6	4.3	1.7	2.8	65	R3	B.1	P
4	4	2	2	2.8	65	R3	B.2	TD
5	5	2.8	2.2	2.8	19	R1	B.3	TD
6	6	4.2	1.8	2.9	65	R3	A.2	P
4	4	1.9	2.1	2.9	13	R1	A.1	TD
5	5	2.8	2.2	2.9	35	R1	C.3	TD
8	8	5.7	2.3	2.9	4	R1	A.4	LR
5	5	2.7	2.3	2.9	4	R1	A.4	TD
5	5	2.7	2.3	2.9	19	R4	B.3	TD

Cat	Score	Exp.	Resd	StRe	Examinee	Rater	Topic	Criterion
5	5	2.7	2.3	2.9	38	R2	B.3	TD
5	5	2.7	2.3	2.9	64	R4	B.4	TD
9	9	6.5	2.5	3	8	R4	A.5	P
6	6	4.2	1.8	3.1	13	R3	C.2	P
9	9	6.5	2.5	3.1	8	R4	C.5	P
6	6	4.1	1.9	3.3	65	R3	B.2	P
8	8	5.8	2.2	3.4	73	R3	A.2	P
5	5	2.4	2.6	3.4	61	R1	A.3	TD
5	5	2.2	2.8	3.7	71	R1	B.2	TD
8	8	5.1	2.9	3.7	9	R2	A.2	P
5	5	1.9	3.1	4.2	66	R1	A.4	TD

Appendix T. Multiple Regression (Assumptions)

Assumption of Multicollinearity

The assumption of no multicollinearity in the data was checked by examining the VIF and tolerance statistics provided in the SPSS output (see table below). There are no VIFs greater than 10. The average VIF is calculated as follows and the value 1.129 is quite close to 1. None of the tolerance values indicate any problems. It is therefore safe to assume that the assumption of no multicollinearity has been met.

$$VIF = \frac{\sum_{i=1}^k VIF_i}{k} = \frac{1.025 + 1.124 + 1.204 + 1.287 + 1.008}{5} = 1.129$$

Model	Collinearity Statistics	
	Tolerance	VIF
1 (Constant)		
ProficiencyEst	1	1
2 (Constant)		
ProficiencyEst	0.985	1.015
BKEstimates	0.985	1.015
3 (Constant)		
ProficiencyEst	0.983	1.017
BKEstimates	0.889	1.124
TaskAversusTaskB	0.83	1.204
TaskAversusTaskC	0.777	1.287
4 (Constant)		
ProficiencyEst	0.976	1.025
BKEstimates	0.889	1.124
TaskAversusTaskB	0.83	1.204
TaskAversusTaskC	0.777	1.287
Gender	0.992	1.008

Table: Multiple Regression: Collinearity Statistics

Assumption of Homoscedasticity and Linearity

The assumptions of homoscedasticity and linearity were checked by examining the ZRESID against the ZPRED graph (below), which shows the dots to be relatively evenly dispersed around zero, with no funneling out of data points or curves which means that the assumptions of homoscedasticity and linearity have been met respectively.

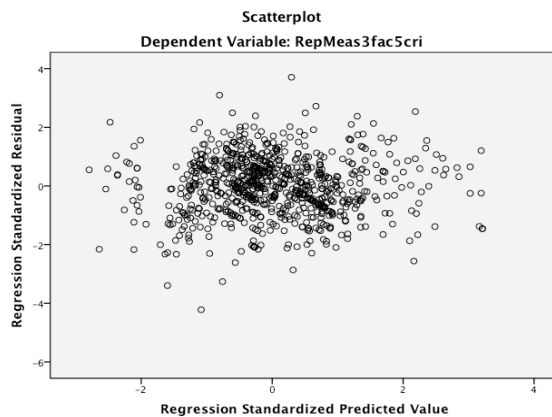


Figure: Multiple Regression: ZRESID against ZPRED graph

Assumption of Normality of Residuals

The assumption of the normality of residuals was checked by examining the histogram and the normal probability plots (see figures below). The histogram looks like a normal distribution with a bell-shaped curve. The normal probability plot also shows a straight line (showing a normal distribution) with the data points representing observed residuals. As can be seen, in a perfectly normally distributed data set, the points all lie on the line which is indeed the case for the data in this study. The assumptions of the model are therefore all met.

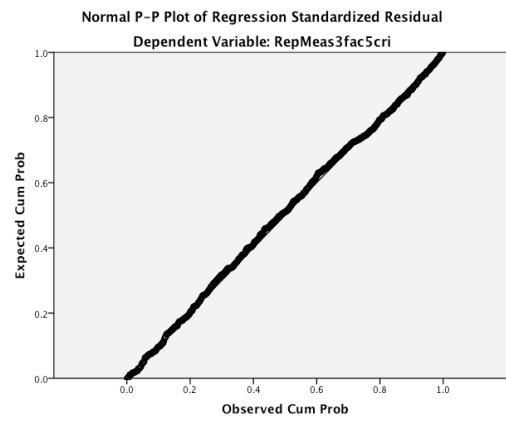
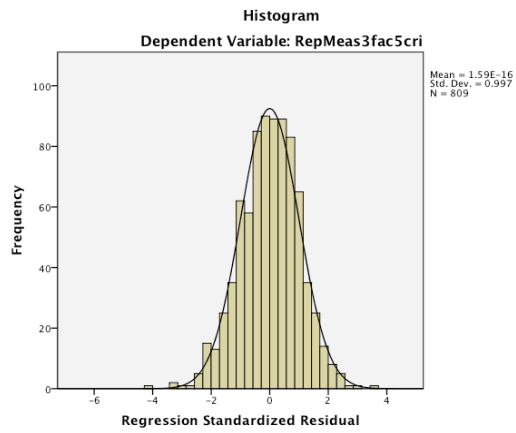


Figure: Multiple Regression: Histogram (left) and Normal Probability Plots (Right)