On Topic Validity in Speaking Tests

Also in this series:

IELTS Washback in Context: Preparation for academic writing in higher education
Anthony Green

Examining Writing: Research and practice in assessing second language writing
Stuart D Shaw and Cyril J Weir

Aligning Tests with the CEFR: Reflections on using the Council of Europe's draft Manual
Edited by Waldemar Martyniuk

Examining Reading: Research and practice in assessing second language reading
Hanan Khalifa and Cyril J Weir

Examining Speaking: Research and practice in assessing second language speaking
Edited by Lynda Taylor

IELTS Collected Papers 2: Research in reading and listening assessment
Edited by Lynda Taylor and Cyril J Weir

Examining Listening: Research and practice in assessing second language listening
Edited by Ardeshir Geranpayeh and Lynda Taylor

Measured Constructs: A history of Cambridge English language examinations 1913–2012
Cyril J Weir, Ivana Vidaković, Evelina D Galaczi

Cambridge English Exams – The First Hundred Years: A history of English language assessment from the University of Cambridge 1913–2013
Roger Hawkey and Michael Milanovic

Testing Reading Through Summary: Investigating summary completion tasks for assessing reading comprehension ability
Lynda Taylor

Multilingual Frameworks: The construction and use of multilingual proficiency frameworks
Neil Jones

Validating Second Language Reading Examinations: Establishing the validity of the GEPT through alignment with the Common European Framework of Reference
Rachel Yi-fen Wu

Assessing Language Teachers' Professional Skills and Knowledge
Edited by Rosemary Wilson and Monica Poulter

Second Language Assessment and Mixed Methods Research
Edited by Aleidine J Moeller, John W Creswell and Nick Saville

Learning Oriented Assessment: A systemic approach
Neil Jones and Nick Saville

Advancing the Field of Language Assessment: Papers from TIRF doctoral dissertation grantees
Edited by MaryAnn Christison and Nick Saville

Examining Young Learners: Research and practice in assessing the English of school-age learners
Szilvia Papp and Shelagh Rixon

Second Language Assessment and Action Research
Edited by Anne Burns and Hanan Khalifa

Applying the Socio-cognitive Framework to the BioMedical Admissions Test: Insights from Language Assessment
Edited by Kevin Y F Cheung, Sarah McElwee and Joanne Emery

Lessons and Legacy: A Tribute to Professor Cyril J Weir (1950–2018)
Edited by Lynda Taylor and Nick Saville

Research and Practice in Assessing Academic Reading: The Case of IELTS
Cyril J Weir and Sathena Chan

Language Test Validation in a Digital Age
Edited by Guoxing Yu and Jing Xu

On Topic Validity in Speaking Tests

Nahal Khabbazbashi
CRELLA, University of Bedfordshire

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781009102490

First published 2021

20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

Printed in XXXX by XXXX

A catalogue record for this publication is available from the British Library

ISBN 978-1-009-10249-0

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate. Information regarding prices, travel timetables, and other factual information given in this work is correct at the time of first printing but Cambridge University Press does not guarantee the accuracy of such information thereafter.

For my family with all the love in the world

تقدیم به خانواده ام با یک دنیا عشق

Contents

Acknowledgements

This book would not have seen the light of day without the love and support of many wonderful people. I would like to take this opportunity to express my gratitude for all they have done.

Much of the research in this volume stems from my PhD thesis and I would be remiss not to thank all those who guided me and helped shape my ideas. I am particularly grateful to Robert Vanderplank, a true gentleman, for supervising my thesis and for his patience and encouragement, David Andrich for teaching me all I know about Rasch analysis, and Ernesto Macaro and Barry O'Sullivan for examining my thesis. I am indebted to my participants for their time and collaboration; this work simply would not have been possible without them. Special thanks go to friends, colleagues, and staff at the Department of Education, University of Oxford and Worcester College for their support during my studies.

Evelina Galaczi has been a mentor, colleague, and the loveliest of friends for over a decade. She has shown me how it is possible to lead with grace, integrity, and kindness. Thank you for being like an older sister to me and for always cheering me on.

My years at Cambridge Assessment English were instrumental in helping me to understand the complexities and challenges of large-scale assessment. I am grateful to friends and colleagues there for generously sharing their professional expertise with me, giving me the opportunity to work on many different and exciting projects, and helping me to become a better researcher.

I was honoured to know and work with the late Professor Cyril Weir. At my first LTF conference, he took time out of his busy schedule to offer feedback on my PhD. Later on, he took me under his wing and introduced me to the CRELLA family. He is deeply missed and I will never forget his kindness. My colleagues at CRELLA have provided me with a wealth of support, knowledge, and experience from the first day I joined. You are a dream team to work with and I learn from you every day.

This project was greatly facilitated by the guidance of Nick Saville and Lynda Taylor as editors for the Studies in Language Testing series. I cannot thank Lynda enough for her wisdom, discerning observations, and critical feedback, which have made the book so much stronger. I am also grateful to John Savage for his editorial assistance and attentive proofreading.

During my PhD and the writing of this book, my friends and family have been a continuous source of support. Thank you to Annie, Bipana, Daniel,

Daniela, Gad, Janeth, Kasia, Laura, Matthias, Megan, Michiel, Mona, Natasa, and Tania for always sharing the good, the bad, *and* the ugly.

Family . . . where life begins and love never ends. My parents have sacrificed so much for me to be where I am today and I will be forever thankful to them. Momzooki, you are an inspiration; thank you for always believing in me and encouraging me to do my best. Babaee, you are the kindest and most gentle of souls; thank you for always giving me the strength to carry on. Nima, Noonoo, Raman, Maman Nahid, Fafar, Kaveh, and little Zani, thank you for always being by my side even from thousands of miles away, for cheering me up with your phone calls, and boosting my confidence when I'm feeling low. Thanks also to Julie, Martin, Melissa, Richard and Grandma Sylvia for welcoming me into the Cole family with open arms. Ammeha, Baba Madani, and Madarbozorg – you are no longer with me but I think of you often and cherish our memories. Finally, I would like to thank Ross for his endless love (and supply of tea while I was writing), for wanting to know me even at my most vulnerable, and for adding a touch of the humanities by suggesting the 'on' in the title of this book.

Series Editors' note

Since the Studies of Language Testing (SiLT) series first appeared in 1995, a key underlying aim of the series has been to publish high-quality doctoral dissertations in order to enable the language testing community to access and benefit from research that makes a contribution to the field but which might not otherwise reach publication. PhDs are selected for inclusion in the series according to a rigorous set of criteria which include:

- being a contribution to knowledge
- being previously unpublished
- having a sound theoretical basis
- being well-referenced to the literature
- being research-based
- being executed with care and thoroughness
- demonstrating analysis and interpretation which is well-founded
- having the style of an academic monograph.

Over the past quarter of a century, over 10 such PhDs have been published in the series – constituting roughly 20% of the total list of over 50 volumes. Among other topics, they report research relating to analyses of test taker characteristics and performance, the testing of reading comprehension, the investigation of washback and impact, and young learner assessment. Many of the authors who were invited to publish their doctoral dissertation as a SiLT volume in the early years of their career have since gone on to become internationally renowned figures in the world of language testing and assessment, including Antony Kunnan, Jim Purpura, Liying Cheng and Anthony Green.

This latest volume to join the PhD subset in the SiLT series is by Nahal Khabbazbashi and it makes an important contribution to the body of literature on the assessment of L2 speaking, specifically the effects of topic and background knowledge on test taker performance. As the author points out, topics are commonly used as a key speech elicitation method in performance-based assessments of spoken language. Nevertheless, the validity and fairness issues surrounding topics have been surprisingly under-researched. Potential research questions focus on the extent to which different topics can be said to be 'equivalent' or 'parallel'. Is it possible that some topics bias against or favour individuals or groups of individuals?

How far does background knowledge of topics have an impact on spoken performance? What are the validity and fairness implications of a potential topic effect due to background knowledge when designing spoken language assessments?

The volume reports a doctoral research study to address these questions, drawing on original data as well as insights from recent empirical and theoretical research. Like several of the previously published PhDs in the SiLT series (e.g. Clapham 1996, Green 2007), the research is grounded in the real-world assessment context of one of the most well-known international English language tests, the International English Language Testing System (IELTS). The volume starts from an up-to-date review of the theoretical and empirical literature related to topic and background knowledge effects on second language performance. This is followed by an accessible and systematic description of a mixed methods research study with explanations of design, analysis, and interpretation considerations at every stage. The conclusion presents a comprehensive and coherent approach for building a validity argument in a given assessment context, and argues for an expansion of current definitions of the speaking construct by emphasising the role of *content of speech* as an important – yet often neglected – feature in speaking assessment. The volume therefore contributes to recent critiques of contemporary models of communicative competence with an over-reliance on linguistic features at the expense of more complex features of communication.

As with earlier PhDs published in the SiLT series, this latest volume should provide valuable source material for postgraduate students and those with an academic interest in language testing and assessment. It will also be a useful resource for practitioners and those working professionally in the field of speaking assessment such as personnel in examination boards, item writers, curriculum developers, and anyone seeking to better understand and improve the fairness and validity of topics used in assessments.

Lynda Taylor and Nick Saville
May 2021

References

Cheng, L (2005) *Changing Language Teaching through Language Testing: A washback study*, Studies in Language Testing volume 21, Cambridge: UCLES/Cambridge University Press.

Clapham, C (1996) *The Development of IELTS: A study of the effect of background knowledge on reading comprehension*, Studies in Language Testing volume 4, Cambridge: UCLES/Cambridge University Press.

Green, A (2007) *IELTS Washback in Context: Preparation for academic writing in higher education*, Studies in Language Testing volume 25, Cambridge: UCLES/Cambridge University Press.

Kunnan, A J (1995) *Test Taker Characteristics and Test Performance: A structural modeling approach*, Studies in Language Testing volume 2, Cambridge: UCLES/Cambridge University Press.

Purpura, J E (1999) *Learner Strategy Use and Performance on Language Tests: A structural equation modeling approach*, Studies in Language Testing volume 8, Cambridge: UCLES/Cambridge University Press.

List of abbreviations

Adj	Adjustment
ANOVA	Analysis of Variance
BK	Background Knowledge
CA	Conversation Analysis
CAF	Complexity, Accuracy, and Fluency
CEFR	Common European Framework of Reference for Languages
CET	College English Test
CPC	Category Probability Curve
df	Degrees of Freedom
DIF	Differential Item Functioning
DTM	Differences That Matter
EAP	English for Academic Purposes
EFL	English as a Foreign Language
ESL	English as a Second Language
ETS	Educational Testing Service
FC	Fluency and Coherence
GA	Grammatical Range and Accuracy
GEPT	General English Proficiency Test
GESE	Graded Examinations in Spoken English
ICC	Item Characteristic Curve
IELTS	International English Language Testing System
IRT	Item Response Theory
ISE	Integrated Skills in English
IST	IELTS Speaking Test
ITA	International Teaching Assistant
L1	First Language
L2	Second Language
L-measure	Lexile Measure
LR	Lexical Resource
LSP	Language for Specific Purposes
LT	Language Testing
M	Mean
MFRM	Many-Facet Rasch Measurement
MnSq	Mean Square
Obs-Exp	Observed minus Expected
OET	Occupational English Test

P	Pronunciation
PSI	Person Separation Index
RMSE	Root Mean Square Standard Error
RQ	Research Question
SCF	Socio-cognitive Framework
SD	Standard Deviation
SE	Standard Error
SLA	Second Language Acquisition
TBLA	Task-based Language Assessment
TBM	Topic Boundary Marker
TD	Topic Development
TEAP	Test of English for Academic Purposes
TESOL	Teaching English to Speakers of Other Languages
TestDaF	Test Deutsch als Fremdsprache
T-LAP	Test of Listening for Academic Purposes
TOEFL iBT	Internet-based Test of English as a Foreign Language
TPC	Threshold Probability Curve
TT	Test Taker

1 Variability in speaking assessment and the role of topic

Introduction

Topics are often used as a key speech elicitation method in performance-based assessments of spoken language. They thus constitute an important area for validity enquiry. For instance, are different topics 'equivalent' or 'parallel'? Can some topics bias against or favour individuals or groups of individuals? Does background knowledge of topic have an impact on performance? Might the content of test taker speech affect their scores – and perhaps more importantly, should it?

In performance-based assessments of speaking, a common practice for eliciting speech is to engage test takers with a topic or range of topics. To address these topics, test takers often draw on their topic-related background knowledge (BK), which generally serves as an information base for performance to be built upon (Bachman and Palmer 1996). To illustrate, a test taker might be asked to talk about an important festival, a newspaper article they have read, or a recent holiday. Test takers would then need to draw on their knowledge and experiences of the topics as well as their language skills in order to formulate a response.

In administering different topics to test takers, there is an underlying assumption that (all other things being equal) the speaking tasks are of equivalent levels of difficulty regardless of the choice of topic and can thus be considered 'parallel'. What logically follows is a second assumption that the individuals' degree of topic-related BK does not have a significant influence on their test results. Evidence to the contrary, however, may suggest that a validity threat has been introduced to the test owing to the influence of the construct-irrelevant factor of BK. 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 BK (Jennings, Fox, Graves and Shohamy 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. A review of the literature on the effects of topic and BK of topic on performance, however, points to a need for more empirical research on these issues, particularly in the context of speaking.

This volume reports on an empirical research study investigating the role of topic and BK of topic in the Speaking module of IELTS (International English Language Testing System); an established and widely used, face-to-face second language (L2) speaking test. It draws on original data as well as insights from empirical and theoretical research to address some of the questions and issues raised so far. By grounding the research in the real-world assessment context of IELTS, this volume allows for an exploration of topic validity against the backdrop of one of the world's most high-stakes English language tests.

Variability and spoken performance

Variability in L2 spoken performance assessment is an area of significant interest and debate from both a theoretical and an empirical standpoint. A number of factors other than the speaking ability in question may have the potential to influence performance in an assessment context (McNamara 1996, Milanovic and Saville (Eds) 1996).

McNamara's (1996:86) model of proficiency, for example, illustrates the complexities involved in performance assessment, while Kunnan (1995:6) 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 L2 performance (see also O'Sullivan 2000 for a synthesis of the literature on test taker characteristics). Other factors that have been identified in the literature as potentially exerting an impact on performance include (but are not limited to): characteristics of the tasks and processing conditions (De Jong and Vercellotti 2016, Luoma 2004, Skehan and Foster 1997); characteristics of the interlocutor(s) and/or rater(s) – such as gender, proficiency level, and personality (Nakatsuhara 2011, O'Sullivan 2000); raters' degree of harshness or leniency when rating (Eckes 2009, McNamara 1996, McNamara, Knoch and Fan 2019, Yan 2014); raters' approaches to scoring and interpretations of rating scale criteria (Baker 2012, Cumming, Kantor and Powers 2002, Lumley 2002, Milanovic, Saville and Shuhong 1996); characteristics of the rating scales (Barkaoui 2007, 2010); degree of acquaintanceship between interlocutors (O'Sullivan 2002); and lastly, socially constructed phenomena such as power relations (Shohamy 2001).

The reason why variability in performance assessment is so critical is that the resulting score from a test is used for making inferences about test takers' abilities and for making important decisions about them. A score, however, can be 'attractively simple' and yet 'deceptive' (McNamara 2000:55). To illustrate, a score of 5 awarded by a harsh rater on a difficult task is meaningfully different from a score of 5 awarded by a lenient rater on an easy task. Extending the argument, we can say that for a cultural topic such as the

Mexican Día de Muertos (Day of the Dead), a score of 5 may have different meanings for a test taker who is familiar with this event from one who has little BK or experience of it.

From a test validity perspective, it is important to monitor the potential effects of factors 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). This now brings us to the parameters of interest in the empirical research reported in this volume: topic – as a *task characteristic* – and BK of topic – as a *test taker characteristic*.

Speaking task characteristics: Focus on topic

Spoken performance, as discussed in the section above, can be influenced by several parameters such as characteristics of the task, the test taker, the interlocutor, the rater, the rating scale and criteria, as well as the interactions between them (McNamara 1996). Speaking test tasks play a pivotal role in assessment; they serve as a link between test takers' underlying abilities and subsequent performance through eliciting samples of speech (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). Of relevance here is a distinction made by Brown, Anderson, Shillcock and Yule (1985) between 'chatting' and 'information-related talk' as representing two ends of a continuum in respect of the purposes of 'talk'. Chatting is viewed as a predominantly social activity that 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' (Luoma 2004:22). These topics are not necessarily discussed in great depth. At the other end of the continuum is 'information-related talk' described as 'speech aimed at transferring information on a particular topic' and is the one more often used in assessment contexts (Luoma 2004:23). 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.

Topic features prominently 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. The topic component carries information in the input of the task and can be 'personal, cultural, academic, or technical' (Bachman and 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). Weir (2005:76) also voices a concern that different topics may elicit ‘responses that are measurably different’. Illustrative examples from the literature include academic and technical topics such as the ‘natural virus’ topic for medicine majors and a ‘computer virus’ topic for computer science majors (Bei 2010), and cultural topics such as the Moon Festival being considered more familiar for Chinese learners compared to St Patrick’s Day (Li et al 2017) (see Chapter 2 for more examples and details of these studies).

Whereas topic is viewed as an important characteristic of tasks in general, it is reasonable to assume that its salience may also be affected by task type. Two task types are of particular relevance here: *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 provided from other sources (Jamieson, Eignor, Grabe and Kunnan 2008). The Internet-based Test of English as a Foreign Language (TOEFL iBT) integrated speaking task, for example, requires the candidate to first read a passage about a campus-related topic, to then listen to a conversation about the same topic, and to subsequently prepare a response that summarises and brings together the information from the two input sources[1].

In contrast, independent 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). An example of a TOEFL iBT independent speaking question is:

> Some people think it is more fun to spend time with friends in restaurants or cafes. Others think it is more fun to spend time with friends at home. Which do you think is better? Explain why.[2]

Unlike the previous integrated speaking task example, the test taker is not supplied with any additional reading or listening input to engage with in addressing this prompt.

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 (Barkaoui, Brooks, Swain and Lapkin 2012, Lee 2006, Luoma 2004).

[1] www.ets.org/toefl/test-takers/ibt/about/content/speaking/q2-integrated-transcript

[2] www.ets.org/toefl/test-takers/ibt/about/content/speaking/q1-independent-transcript

Of particular relevance to this discussion are the different approaches to addressing topic-related BK in these task types. In integrated tasks, an attempt is made to minimise and/or mediate the (negative) impact of BK through the provision of input in the form of reading and listening texts. Independent tasks, on the other hand, require test takers to draw on their own BK. This absence of input in independent tasks has been criticised for not allowing an 'equal footing' (Weigle 2004:30) for test takers who bring varying degrees of BK to a test and for the restriction of topics to 'fairly bland' ones (Brown et al 2005:1). In this light, integrated tasks are viewed as 'promoting equity or fairness' (Huang 2010:4). We can also argue that by providing test takers with the necessary ideas for responding to a topic (instead of asking them to generate ideas and rely on their own BK), 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 BK of topic are likely to be manifested more markedly in independent tasks. Recent research, however, suggests that integrated tasks may not be necessarily 'immune to the influence of prior topical knowledge on scores' and that BK can be a 'significant determinant' of speaking test performance regardless of task type (Huang, Hung and Plakans 2018:43).

Test taker characteristics: Focus on background knowledge of topics

A discussion of task topics is inextricably linked to test takers' BK of topics. BK is referred to in the literature under different terms such as content knowledge, prior knowledge, schematic knowledge, topical knowledge, and world knowledge. These terms are often used interchangeably although there has been a recent move towards establishing the nuances between the different terms (see for example Banerjee 2019 and O'Reilly and Sabatini 2013). Broadly speaking, a facilitative role for BK on performance has been posited in the theoretical literature.

A central role, for example, has been ascribed to BK in language comprehension as formalised in schema theory (Bartlett 1932, Carrell and Eisterhold 1983, Rumelhart 1980).

> [T]ext, any text, either spoken or written, does not by itself carry meaning. Rather according to schema theory, a text only provides directions for listeners or readers as to how they should retrieve or construct meaning from their own previously acquired knowledge. This previously acquired knowledge is called the reader's *background*

> *knowledge*, and the previously acquired knowledge structures are called *schemata* (Carrell and Eisterhold 1983:556; emphases in original).

A facilitative role for topic familiarity on performance has also been suggested in Skehan's (1998) framework of task processing conditions. It is hypothesised that the more familiar a topic is to an individual, the less cognitive load it poses 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). In a similar vein, Robinson's triadic componential framework (2001) views prior knowledge as a cognitive complexity dimension where unfamiliar tasks and those for which individuals' prior knowledge is lower can increase task complexity, leading to 'a *depletion* of attentional and memory resources' (Robinson 2001:308; emphasis in original) affecting the accuracy and complexity of performance.

In the field of language assessment, Bachman and Palmer (1996:65) define topical knowledge as 'knowledge structures in long-term memory' that can have a substantial effect on performance. Topical knowledge features as one of the five main components of Bachman and Palmer's (1996) model, 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). BK is also often considered in relation to potential sources of test bias where the test task 'contains content or language that is differentially familiar to subgroups of test takers' (O'Sullivan and Green 2011:61).

Despite the pronounced role attributed to BK of topics on performance from a theoretical standpoint, the results of empirical studies on the subject are often mixed and inconclusive (see Chapter 2). One possible reason is the various ways in which BK has been operationalised in the literature, for example, as knowledge related to academic field of study (Clapham 1996), cultural background (He and Shi 2012), gender (Lumley and O'Sullivan 2005), religious background (Markham and Latham 1987), and personal interest in topics (Jennings et al 1999). Furthermore, the majority of empirical studies have explored the effects of BK on reading and listening comprehension with fewer studies focusing on the performance skills.

Only a handful of studies have exclusively examined topic and BK 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 BK for the spontaneous generation of ideas necessary for addressing independent topic-based tasks. Given the importance attributed to the two associated factors of topic and BK of topic, a closer examination of these two variables on speaking performance is warranted.

This is particularly so in assessment settings where topic plays a critical role in eliciting speech, as is the case in the IELTS Speaking test. In the next sections, I will provide more details of the research context and illustrate the centrality of topics in the test as the main 'driver' of speech.

Research context

What is IELTS?

Research in language assessment, more often than not, is linked to specific exams, testing instruments, and validation efforts. By grounding research in real-world assessment contexts, results of studies can influence and shape testing practices with the potential to impact a large number of individuals and organisations, particularly in the case of large-scale standardised tests.

IELTS is one of the world's most popular English language tests used for study, migration, or work. It has a candidature of over 3.5 million per year and is taken in 1,600 test centres in more than 140 countries around the world[3].

IELTS has two modules: Academic and General Training. The Academic module is designed to assess English language proficiency for those applying for higher education or professional registration. General Training is used to measure proficiency for more practical use in social contexts and is used for migration as well as other purposes such as training, secondary education, and work in English-speaking environments[4]. Both modules have four papers covering the skills of listening, reading, writing, and speaking. The Listening and Speaking papers are common across the modules with the subject matter of the Reading and Writing sections as the main differentiating factor between the two.

IELTS scores are reported on a nine-band scale from non-user (a score of 1) to expert user (a score of 9). There is no pass or fail 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'[5]. Some guidance on overall IELTS scores based on the linguistic demands of academic curricula is provided though this is not designed to be prescriptive and instead, organisations and institutions are encouraged to set their own minimum scores on the basis of their specific requirements[6].

[3] www.ielts.org

[4] www.ielts.org/about-ielts/ielts-test-types

[5] www.ielts.org/-/media/pdfs/ielts-test-score-guidance.ashx

[6] For more information on the test as well as the latest research on IELTS, visit the IELTS website (www.ielts.org). For information about the historical development of IELTS, the interested reader is referred to Davies (2008). The volume chronicles the evolution of IELTS against the historical backdrop of academic English language proficiency testing within

IELTS can be safely 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). Putting test takers at the heart of assessment, Shohamy (2001:102) defines the criterion for a high-stakes test as ‘whether the results of the test lead to detrimental effects for the test takers’. 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 (LT) community (Fulcher and Davidson 2007). It is this very high-stakes nature of IELTS that demands rigorous research on every aspect of the test in order to ensure its validity and fairness.

The IELTS Speaking test

The IELTS Speaking test (IST) is a face-to-face oral interview between a test taker and a certified IELTS examiner[7]. The interview lasts between 11 and 14 minutes and is recorded. There are three main parts in the IST. Following an introduction, Part 1 is 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, the *Individual Long Turn*, the candidate is presented with a printed task card, which requires an extended talk on a specific topic for 1 to 2 minutes. The task card includes points that the candidates can cover in their monologue. Prior to the monologue, the candidate is given 1 minute of silent planning time as well as pencil and paper for making notes. The third part of the test, which lasts about 4 to 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 final part aims to provide the candidate with an opportunity to discuss these more abstract themes and topics.

Part 1 consists of three multi-question topic sets (or *topic frames*). The first frame involves a general topic such as work or studies and the remaining two sets involve other familiar topics. Part 2 consists of one topic-based monologue task. Similar to the Interview task, the Two-way Discussion often consists of two topic sets. The topic sets for Parts 1 and 2 are randomly

the UK higher education system. In his reflections, 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. This is a point worth emphasising, as practicality concerns – particularly those associated with large-scale assessment – are often neglected in research studies.

[7] For a comprehensive historical overview of the IST see Nakatsuhara (2018).

assigned to test takers and the topic sets from Part 3 of the test are thematically linked to the Part 2 topic.

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**' (IELTS Examiner Training Material cited in Seedhouse and Harris 2011:72; emphasis in original) 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 construct underlying IST 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'[8].

The IELTS speaking scale is a nine-band analytic scale consisting of four criteria: Fluency and Coherence, Lexical Resource, Grammatical Range and Accuracy, and Pronunciation. 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 IST is single-marked and test reliability is assured through examiner training and certification, standardised procedures, a sample monitoring system, a 'jagged profile' system, and routine validation of task, candidate and examiner performance (Taylor 2007a:29).

Topic as main 'driver' of speech in the IST

Let us now consider the IST and its structure with a view to contextualising previous discussions on topic effects on performance with reference to specific test features.

In all three parts of the IST, speech is elicited by means of different topics, from more familiar ones in Part 1 to more abstract themes in Part 3. This gradation is designed to cater to the IELTS candidature who can vary widely in their speaking ability levels. The familiar/unfamiliar and concrete/abstract continua are used to span this divide in terms of demands on candidates and scope for sufficient production to allow for meaningful evaluation.

[8] www.ielts.org/-/media/publications/information-for-candidates/ielts-information-for-candidates-english-uk.ashx

Based on their findings from a conversation analysis of transcribed IELTS spoken performances, Seedhouse and Harris (2011) identify 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' (Seedhouse and Harris 2011:83). 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. Drawing on this research amongst others, Seedhouse (2018) views topic as a fundamental construct within the IST exhibiting what he calls 'a dual personality' (2018:114): 'topic-as-script' and 'topic-as-action'. The former refers to 'the homogenised topic which examiners give to candidates' and the latter refers to the 'diverse ways in which candidates talk a topic into being' (2018:114) which could subsequently impact performance scores.

Topic, therefore, constitutes the main vehicle for driving talk in the IST and yet, a specific topic development or content-oriented criterion is surprisingly absent in the IST Band Descriptors, which might explain why test takers may not always elaborate on topics (Seedhouse and Harris 2011). Topic is referred to under the *Fluency and Coherence* scale to differentiate higher proficiency levels: Band 8 – 'develops topics coherently and appropriately' and Band 9 – 'develops topics fully and appropriately', with little information on the distinctions between 'coherently' and 'fully'. For lower levels, topic is referred to under the *Lexical Resource* scale to differentiate extent of lexical knowledge on familiar and unfamiliar topics: Band 3 – 'has insufficient vocabulary for less familiar topics' and Band 3 – 'able to talk about familiar topics but can only convey basic meaning on unfamiliar topics' (see www.ielts.org/-/media/pdfs/speaking-band-descriptors.ashx?la=en for a public version of the Band Descriptors). What is left unsaid, of course, is how familiarity is determined for different candidates.

A topic or content-oriented criterion capturing the extent to which ideas are developed has not been explicitly defined as part of the construct of IST and other speaking tests more widely (Elder, McNamara, Kim, Pill and Sato 2017, Sato 2012). The inadequacy of the current IELTS speaking rating scale in dealing with 'off-topic responses' or in addressing limited topic development was commented on by the IELTS examiners surveyed in an early study by Brown and Taylor (2006) and more recently in the research conducted by Inoue, Khabbazbashi, Lam and Nakatsuhara (2021). Both studies recommended the inclusion of a criterion focusing on task response/ topic development. There are perhaps several practical reasons why this recommendation has not been applied to the IST such as increased cognitive demand on raters and the necessity to award separate scores for each test part. Nevertheless, the role of topic development in allowing the test to achieve its main purpose of generating samples of rateable speech requires careful consideration regarding whether a content-oriented criterion – which explicitly emphasises the development of ideas and topics – should be included in the scales (Elder et al 2017, Sato 2012).

There are two additional features of the test that can further amplify the topic effect. Firstly, all speaking tasks on the IST are independent and not integrated and thus rely on test takers bringing their own BK in addressing the task topics. While the multi-task structure of the test, the inclusion of a mix of familiar/concrete and unfamiliar/abstract topics, and provision of scaffolding in the form of planning time and prompts are designed to obviate or minimise the topic effect, test takers nevertheless need to draw on their BK and deal with the spontaneous demands of addressing different topics online. 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' (2011:15). Seedhouse and Harris (2011) identify the dominance of the examiner script as being aligned to the institutional goals of achieving reliability and standardisation. I would like to argue, however, that the absence of choice for test takers combined with unequal power dynamics which, taken together, do not allow candidates to shift or navigate topics can in fact compromise the main institutional goal of eliciting rateable samples of speech when test takers do not have the necessary BK to address the topics or tasks.

Research focus and validation framework

What I have tried to illustrate so far is the important role played by topics and BK of topics in speaking performance assessment: topics, in providing a temporary meaningful context for the elicitation of rateable samples of speech; and BK of topic in providing the information base on which speech is built (Bachman and Palmer 1996). More research on the influence of these variables on speaking performance is needed, particularly in the IST where, as demonstrated, the topic effect is likely to be more prominent with the potential to introduce a validity threat to the test (Messick 1989). This volume, in response, focuses on an examination of these two critical parameters and their effects on speaking performance in the specific assessment context of the IST. The research will be guided by Weir's (2005) socio-cognitive framework for language test validation. This framework provides a practical, systematic, and coherent approach to validation activities by bringing together social, cognitive, and evaluative dimensions of language use and linking them to the context and consequences of test use. In the next section, I will provide a brief description of the framework before outlining how it will be applied to my research in building a topic validity argument.

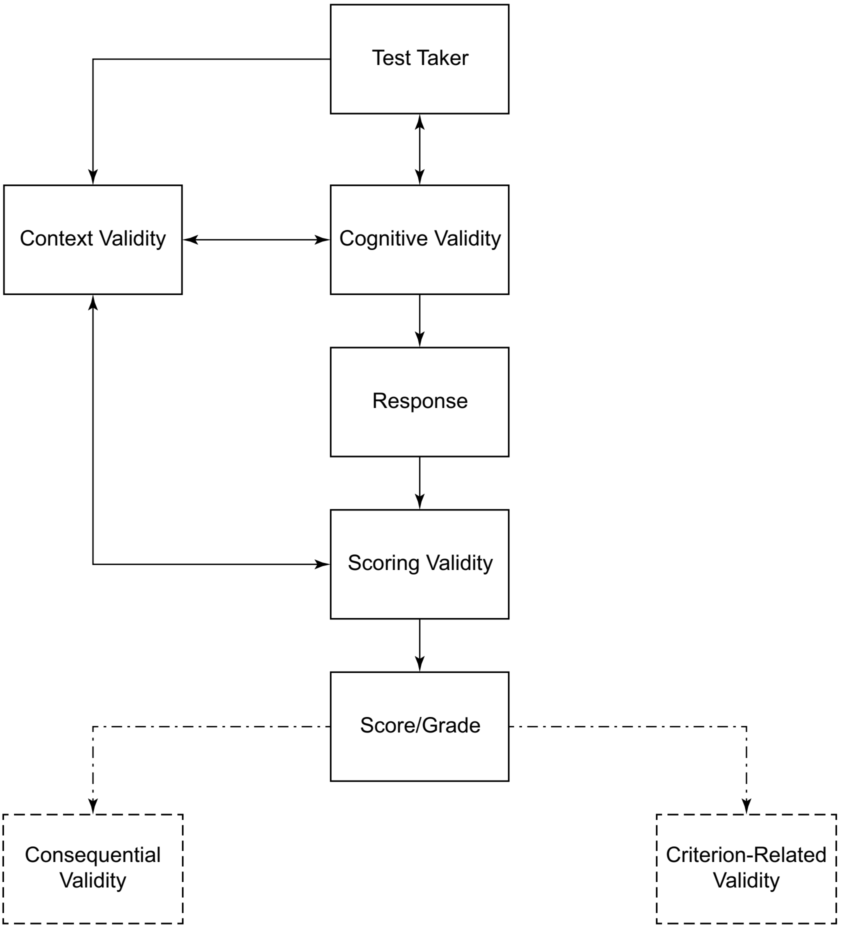
The socio-cognitive framework for test development and validation

The socio-cognitive framework (SCF) (Weir 2005, later modified in O'Sullivan and Weir 2011) consists of six central elements: test taker characteristics, cognitive validity, context validity, scoring validity, consequential validity, and criterion-related validity (see Figure 1.1). 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 the kind of validity evidence that 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'.

As its title suggests, the SCF conceptualises the language ability construct as lying in the interaction between both *cognitive* elements and *contextual* features (depicted visually in the bi-directional arrows between the context validity and the cognitive validity components of the model). Importantly, the model emphasises the centrality of the test taker and their characteristics as exerting a direct influence on the way the contextual parameters of the model are cognitively processed (Weir 2005). The arrows represent the interactions between characteristics of the test taker, features of the task, and the mental processes that have been activated by the task parameters, leading to a response that is subsequently scored. These cognitive, contextual (social), and evaluative (scoring) components constitute the 'core' elements of the framework (O'Sullivan and Weir 2011:24) with a 'symbiotic' relationship between them.

Messick's (1989) concern for a consideration of test consequences is reflected in the consequential validity component of the SCF which, along with criterion-related validity, is located at the bottom of the graphic/ figure and after a score has been generated. Temporally speaking, these two types of validity evidence can be generated once a test has been developed

Figure 1.1 Socio-cognitive framework for test development and validation (O'Sullivan and Weir 2011:21)



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. This stance can however be challenged: if a new test is being developed or an existing test is being re-engineered to fit within a wider framework such as the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001), then perhaps evidence of criterion-related validity may be necessary at an earlier stage. In a similar vein, a consideration of consequential validity from the outset of the test design process – rather than a *post hoc* activity – can help with positive impact by design[9].

By conceptualising validity as 'multifaceted', different sources of validity evidence (context, cognitive, scoring, consequential, and criterion-related) are fitted together in the SCF to present a unified approach to validity. The comprehensive guidance provided for different approaches and methods in the systematic collection of validity evidence at different stages of the test cycle makes this framework not only informative from a theoretical perspective but also practical to operationalise. This is one of the strengths of the SCF and is evidenced in the variety of projects and assessment settings where the framework was used for the development, revision, and validation of language tests[10].

Applying the SCF to research focus

In this section, I will first describe the different elements of SCF (Weir 2005) in more detail before explaining how the two parameters of interest in my research – topic and BK of topic – will be systematically explored within the framework with illustrative research questions (RQs).

The *test taker* element of the framework is directly linked to *cognitive validity*, together representing the individual candidate in a test. Drawing on the work of O'Sullivan (2000), test taker characteristics are divided by Weir (2005) into three groups – physical, physiological, and experiential – with different features within each group hypothesised to be in constant interaction with one another. Physical/physiological characteristics refer to 'fixed' biological features such as age or sex 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 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. BK (or general world knowledge/topic knowledge) is subsumed under experiential characteristics in the SCF and viewed as an important feature likely to influence test performance. As

[9] I am grateful to Lynda Taylor for bringing my attention to this point.

[10] Examples include the College English Test (CET) and the Test for English Majors in China, a range of examinations provided by Cambridge Assessment English (see Cheung, McElwee and Emery (Eds) 2017, Geranpayeh and Taylor (Eds) 2013, Khalifa and Weir 2009, Shaw and Weir 2007, Taylor (Ed) 2011), the Graded Examinations in Spoken English (GESE) and the Integrated Skills in English (ISE) by Trinity College London, the General English Proficiency Test (GEPT) by the Language Training and Testing Center in Taiwan, and the Test of English for Academic Purposes (TEAP) in Japan.

O'Sullivan and Green (2011:40) argue, individuals from different age or cultural backgrounds are 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 therefore important to demonstrate that those characteristics that are irrelevant to the construct under examination do not exert 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 1994:1–2). In Weir (2005), DIF was originally 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 'BK of topic' is regarded as an experiential test taker characteristic, then an important piece of topic validity evidence is to demonstrate that no individuals or groups of individuals have been unfairly biased against as a result of task topics. A possible RQ that can guide the collection of validity evidence in relation to the test taker element of the SCF can be formulated as follows: *Will (any) differences in test takers' levels of background knowledge of topics have an impact on performance?*

The *context validity* component of the SCF is used as a superordinate category related to tasks and the performance conditions under which they are performed. Within test constraints, Weir (2005:56) suggests that performance conditions should be as similar to authentic language use as possible. Context validity includes characteristics of the task setting, the demands of the task both in terms of linguistic input and output, characteristics of the interlocutor, as well as administrative settings of the test event.

As a feature of task input, a significant role is attributed to topic in the SCF and Weir (2005) cautions that different topics may elicit 'responses that are measurably different' (2005:76). A possible RQ that can guide the collection of context validity evidence can be formulated as follows: *To what extent are the different topics used in parallel versions of a task similar in terms of difficulty measures?*

The *cognitive validity* component relates to the mental processes activated by tasks and should reflect an established theory of the mental processes underlying the construct of interest. For a speaking test, for example, cognitive validity evidence should 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 generalise to' (Weir, Vidaković and Galaczi 2013). Weir's (2005) cognitive validity component for speaking is comprised of executive processes, executive resources, and monitoring features in line with Levelt's (1989) widely recognised model of first language (L1) speech. The 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. I would like to draw attention, however, to the 'conceptualisation' stage of speech processing (Levelt 1989), which is concerned with the generation of ideas, amongst others, and is thus the most relevant aspect of the cognitive validity component to the parameters under examination in my research. The generation of ideas is viewed as a cognitively demanding process increased in line with the increased 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, along which tasks can vary in terms of the cognitive demand they pose on the test taker (Field 2011, Weir et al 2013).

Earlier in the chapter, I touched on the format of the IST and how speaking tasks are designed to increase in difficulty as a function of familiarity and abstractedness from more familiar and concrete topics in Part 1 to less familiar and abstract topics in Part 3. A related RQ that can guide the collection of cognitive validity evidence can be formulated as follows: *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 in terms of speaking includes considerations of rating criteria, rating scales, and the application of the rating scales by human raters (or machines in the case of automated assessment). Citing Shaw and Weir (2007:143), Taylor (2011:29) interprets scoring validity for speaking as follows:

> 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 BK of topics are not influenced by systematic rater tendencies.

Consequential validity in the SCF is also a superordinate term referring to the impact of tests on institutions and society, washback in the classroom or workplace, and test bias. As previously noted, however, Taylor (2011) argues for a consideration of test bias to be subsumed under test taker characteristics. For this reason and due to the scope of the research, other aspects of consequential validity are not addressed in this volume and evidence related to test bias is collected in relation to the test taker element of the SCF.

Lastly, in relation to *criterion-related validity*, Weir (2005) invites us to consider evidence external to the test that lends further support to 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) that can lend support to 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 common example is the aligning of language tests to the CEFR (see Bachman 2011 for a critical review of issues related to linking practices 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).

The third category of evidence for the equivalence of different forms relates directly to the focus of the research in this volume. If different forms of a speaking test consist of different combinations of topics, then a critical piece of criterion-related validity evidence can be derived from establishing topic comparability. A related RQ that can guide the collection of criterion-related validity evidence can be formulated as follows: *To what extent are parallel forms of a language proficiency interview (consisting of different topics) comparable in terms of difficulty?*

Note that numerous combinations of questions can be posed under different validity components of the SCF, which can in turn be addressed using a variety of approaches in collecting validity evidence. This is a strength of the framework in allowing for an examination of any number of hypothetical interactions between various features and components. Correspondingly, different kinds of evidence can be gathered to answer each RQ. The types of evidence will reflect the nature of the questions posed and the scope of validation efforts will reflect the scope and limitations of single projects. In the next chapter, I will draw on a review of the literature to formulate and refine the RQs that can help generate topic validity evidence for the IST as guided by the SCF and within the practical restraints of the project.

Structure of the volume

In this volume, I will be examining the effects of two critical parameters – topics and BK of topics – on speaking performance in the specific assessment context of the IST. Guided by the SCF (Weir 2005), I will draw on insights from relevant literature as well as original empirical data to build a topic validity argument for the IST and to contribute to current knowledge in the related fields of applied linguistics and LT.

My aim in this introductory chapter was to identify the two variables of interest in the research, familiarise the reader with the specific assessment context of the IST, set the scene by highlighting the critical role of topic and BK within the IST, and provide an overview of the validation framework guiding the study. The rest of the volume is organised as follows:

Chapter 2 provides a state-of-the art review of the theoretical and empirical literature related to the parameters of interest drawing from theories of speech production, task-based approaches to language learning and instruction, validation frameworks and empirical research across all four skills as well as IELTS-specific research. The chapter concludes by summarising the methodological implications of the reviewed literature.

Chapter 3 considers the complexities involved in performance assessment and how different factors such as raters, task types, criteria, and the interactions between them are at play in contributing to score assignment and reporting. Following a review of different psychometric approaches to measurement, the chapter introduces Many-Facet Rasch Measurement, explicating how it can allow for an examination of the influence of the main factors (or facets) of interest i.e. topics and BK of topics, independent of but adjusted for the influence of other facets. The conceptual-psychometric framework by Eckes (2009) is subsequently discussed with details of how it was adapted to the research in order to guide the systematic collection and analysis of data.

Chapter 4 opens with the RQs guiding the study and provides a detailed account of the methodology including the research setting, participants, instruments used at different stages of data collection (e.g. speaking tasks, BK questionnaires, and C-tests with test taker participants and rating scales, language functions checklist, and interviews with rater participants), data collection procedures and analyses. These are complemented with explanations for design, analysis, and interpretation considerations at every stage.

Chapters 5 and **6** examine the topic effect from multiple viewpoints including the perspectives of scores, raters, and test takers while considering the influence of other parameters of interest such as BK, task types, and proficiency levels. The effects of topics on task difficulty, how raters award scores and make decisions, and the kinds of strategies that test takers adopt when faced with problematic topics are discussed while also highlighting issues related to local validity of international tests, test fairness, and bias.

Chapter 7 weaves together and evaluates the different strands of research findings in building a topic validity argument. The chapter concludes with implications and recommendations for the IST as well as speaking tests more broadly, arguing for a revisiting of the notion of task 'difficulty' in tests, increasing examiner flexibility and test taker agency, moving towards integrated assessment, and finally expanding the current definitions of the speaking construct by emphasising and including content of speech as part of the test construct.

2 Insights from multiple research domains

In this chapter, I will provide a critical review of different strands of scholarly research related to the parameters of interest, namely topic and BK of topic, drawing on: insights from theories of speech production; task-based approaches to language learning; test validation frameworks; as well as empirical research across the four skills of reading, listening, writing, and speaking. I will also discuss findings from studies that are directly related to the IST. The chapter will end with an overview of implications of the literature review for the research design of the main study.

Insights from second language acquisition research

Topic, as a task-processing condition that 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).

Drawing predominantly on cognitive processing approaches to language learning, Skehan's (1996:52) framework provides a systematic and principled method for 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 that might pose additional pressure on communication: 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.

Topic familiarity, in this framework, features in the cognitive complexity component and is associated with the *conceptualisation* stage of Levelt's (1989) influential model of monolingual speech production in 'generating an idea or set of ideas for expression' (Field 2011:74) and 'accessing relevant aspects of schematic knowledge' (Skehan 1996:52). In Levelt's (1989) model, the speaker first activates different concepts from memory before deciding on the content and the order in which it will be organised. These 'pre-verbal' messages therefore contain the information base for translating meaning into language though at this stage they remain non-linguistic (de Bot 1992).

Skehan (1998, 2001) 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, Xiaoyue, Qian and Wang (2012) also propose another general framework for task-based research in which studies can be categorised depending on whether they focus on *pre-task*, *during-task*, 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 (Nitta and Nakatsuhara 2012, O'Grady 2019, Ortega 1999, Wigglesworth 1997). 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 performed, can be examined (Skehan et al 2012). The post-task stage refers to the potential influence on performance of anticipating a post-task activity that is linked to the task.

From a theoretical perspective, topic familiarity in the Skehan et al (2012) framework is conceptualised as a form of preparedness, linked with the pre-task phase. 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 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 more critical look at the research, however, reveals a rather loose link between structured/unstructured tasks and topic familiarity; task structure in Skehan and Foster (1999:100) is broadly conceptualised in terms of time sequence and predictability inherent in the task, hypothesised to potentially reduce the processing demands of tasks and enhancing 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 storylines. 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' (Skehan and Foster 1999:103). In the 'golf' episode, however, Mr Bean plays crazy golf and, following a bad shot that 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. This interpretation, however, can be questioned, as it assumes a causal link between the concepts of sequential task structure and topic familiarity, which was not independently verified.

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 [a] 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 [a] 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'.

Content familiarity and task repetition were the two factors examined in Qiu's (2020) experimental study with 60 Chinese learners of English. Four monologic tasks – two picture tasks and two short speech tasks – were included in the study. Task topics were designed to vary in familiarity based on (a) learners' cultural background (Chinese wedding gift versus Western wedding gift) and (b) experiences (finding lost items versus job-hunting plans after graduation). Levels of familiarity were established after each task performance; familiarity scores confirmed the familiar/unfamiliar topics as intended by the researcher. Spoken performances were analysed in terms of complexity, accuracy, and fluency (CAF). Findings suggested that participants produced 'structurally more complex oral discourses' (Qiu 2020:756) under the familiar conditions though no significant effects were found for measures of lexical richness, accuracy, or fluency. Bui (2014), on the other hand, reported a positive impact of topic familiarity on accuracy and fluency. Bui and Huang (2018) also examined the impact of topic familiarity on a number of different fluency measures with findings suggesting a positive influence on temporal aspects of fluency such as speech rate and number of mid-clause breakdowns; however, there was little effect on mean length of run or number of fillers and repairs.

Taken together, the body of SLA research suggests that topic familiarity can have a positive impact on performance in relation to inter-language measures of fluency, complexity, and accuracy though results are not always consistent across studies.

Insights from task-based language assessment research

Task-based language assessment (TBLA) is defined by Norris (2016:232) as 'the elicitation and evaluation of language use (across all modalities) for expressing and interpreting meaning, within a well-defined communicative context (and audience), for a clear purpose, toward a valued goal or outcome'. Within TBLA, the notion of *task difficulty* as a function of the cognitive demands and features of tasks is quite appealing; 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, Brown, Hudson and Yoshioka (1998) devised a matrix of task features that could be systematically manipulated 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 were estimated. Norris, Brown, Hudson and Bonk (2002) subsequently implemented the matrix by constructing 13 complex tasks that varied in difficulty as a function of different combinations of cognitive factors and subsequently evaluated task-based performance. Results were not as anticipated; *a priori* difficulty estimates of the tasks failed to predict systematically 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 within an LT context has been questioned by Fulcher (2003), drawing on evidence from several LT studies that 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 such as Fulcher's (1993) comparative study of a text-based interview with picture description and group discussion tasks, and Bachman, Lynch and Mason's (1995) comparative study of two different speaking task types. In both studies, substantial differences in task types only resulted in very small albeit statistically significant differences in scores and task difficulty measures. On the basis of these findings, the authors make a logical inference: should differences in task types fail to result in large effects on performance then by the same token, should task type be held constant, changes in task conditions are unlikely to yield significant results with large effect sizes. The importance of establishing 'practical significance' is then highlighted by Fulcher (2003:65).

The term *practical significance* refers to differences that 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 where 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 particular relevance in LT 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.

Now let us return to the discussion of task difficulty. Bachman (2002) raises a fundamental issue with the conceptualisation of task difficulty as an inherent feature of the task (see also Révész 2014, Norris 2016 for critical reviews on the subject). Instead, he convincingly argues for difficulty to be viewed 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).

Whereas Bachman (2002:466) acknowledges code complexity as a 'unique' characteristic of tasks, he views the remaining two components, i.e.

cognitive complexity and communicative stress, as factors that interact with characteristics of the test taker and therefore not inherent to the task. He poses the question 'wherein lies difficulty?', arguing that it 'resides' in the complex interactions between different features involved in performance assessment (2002:466). He also emphasises the need for methodologies that adequately address such issues in research.

The discussions in this section have three clear implications for the focus of the research reported in this volume: firstly, it is the extent to which topics and BK of topics can have an influence on *test scores* – rather than interlanguage measures – which is a concern; secondly, it is the extent to which (any) statistical significance translates into meaningful practical significance that needs to be established; and thirdly, any adopted methodology needs to take into account the slippery notion of 'difficulty' and adequately model the interactions between different components of assessment.

We now turn our attention to a critical review of studies that have specifically examined the role of topic and BK of topic on L2 performance, classified according to the language skills they relate to: reading comprehension, listening comprehension, writing, and speaking. The purpose of the review is twofold: firstly, to establish whether the importance ascribed to topic and BK of topic in the theoretical literature (e.g. Bachman and Palmer 1996, Skehan 1998, Weir 2005 – see Chapter 1) is echoed in the findings from empirical research. Secondly, to critically reflect on the methodologies adopted in various studies in order to inform the design of the research study addressed in this volume (see Chapter 4).

Insights from reading comprehension research

The results of empirical research on the effects of BK on reading comprehension have been inconsistent and far from conclusive. The majority of this research concentrates on general versus 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 BK on the reading comprehension of international students in pre-sessional courses at university were investigated. BK was not independently measured but was assumed based on 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 also 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 BK 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 BK effect comes from the seminal work of Clapham (1996). In her study, the relative influences of BK 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. It was found that texts varied substantially in their degree of subject-specificity from general to highly specific. BK 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 BK 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 introductions to academic articles versus passages describing research processes in specific fields). Whereas significant variance was explained by both language proficiency and BK, the study showed that BK differentially affected students at different proficiency levels. The intermediate proficiency group of students were most influenced by their BK but the same trend was not observed for low- or high-proficiency students. Clapham (2000:515–516) explains the findings 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.

She went on to conclude 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 that informed the revision of IELTS in eliminating subject-specific reading and writing modules for different academic disciplines (Clapham 2000).

The hypothesis that students may draw on their BK above a certain proficiency threshold (Alderson and Urquhart 1985) was put to the test by Krekeler (2006). This study examined the effect of BK on the reading comprehension of 486 international students at German universities in an ESP assessment context. In addressing the weaknesses of previous studies in measuring BK, Krekeler used a different approach by (a) asking test takers to explain key terminology in the texts before test administration (limited to 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.

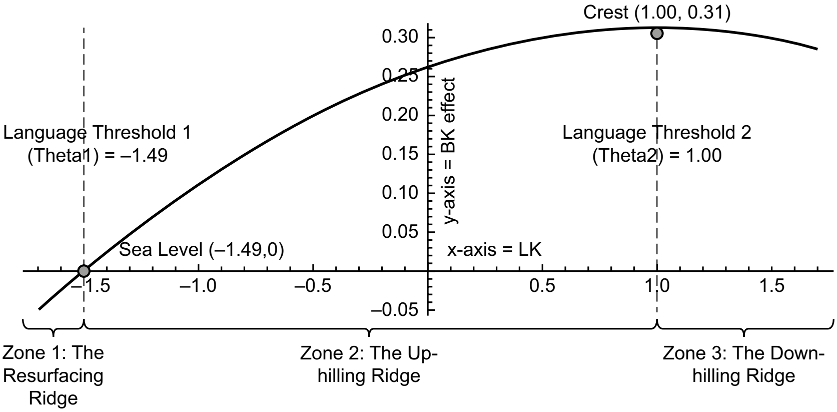
While these measures improved on other assumption-based measures of BK, they are arguably too crude overall to capture nuances in degree of BK. Nevertheless, these measures, particularly the binary yes/no question, can serve as a clear-cut indication of familiarity. Findings from the study showed significant differences between means on the reading test scores of the two groups on all three BK grouping criteria. BK was found to be equally important for both high- and low-proficiency students with the exception of advanced test takers. A closer examination of interaction effects showed little evidence in support of the threshold hypothesis leading the author to suggest that 'if thresholds exist at all, they are fuzzy and may even be chance events' and that 'it would be more sensible to assume that background knowledge, or lack thereof, will usually affect test performance' (Krekeler 2006:123) although 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 BK 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 (matching items, true–false items, open-ended questions, summary tasks, etc.). Prior to the reading tests, participants' level of English and BK 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 10 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 BK 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).

In their review of the literature on BK and reading performance, Cai and Kunnan (2019) point to the limitations of previous studies in adequately measuring BK and the analysis techniques applied. They address these limitations in their own study by using a subject-specific knowledge test as a measure of BK and bifactor-multidimensional item response theory as the analysis technique, respectively. With a large sample size of approximately 1,500 nursing students, this study examined the interaction between BK and language proficiency on reading performance using a nursing knowledge test as a measure of BK, a grammar test as a measure of proficiency, and a nursing English reading test as a measure of Language for Specific Purposes (LSP) reading ability. The analyses revealed an 'up-then-down pattern' of BK on LSP reading as follows: for very low-level proficiency, there was a negative effect of BK; as language proficiency increased (for medium level), there was a positive effect of BK which peaked at a certain level; when proficiency was high enough, the influence of BK became less prominent, 'stepping down from its full potential' (see Figure 2.1).

Figure 2.1 The island ridge curve illustrating the moderation of language knowledge on background knowledge effect in affecting LSP reading performance (Cai and Kunnan 2019:9)



Note: BK = Background knowledge, LK = Language knowledge.

The study's limitations – for example, the use of a grammar test as a measure of language proficiency or restricting the research to the nursing discipline – may make it difficult to extrapolate results to other disciplines. Nevertheless, the study's use of advanced statistical techniques and increased measurement accuracy has enhanced our understanding of the nuanced ways in which BK interacts with language proficiency.

Taghizadeh Vahed and Alavi (2020) examined the impact of discipline-related BK on EAP reading comprehension performance while also taking into account the effects of task type. 206 civil engineering students at three language levels took part in the study. Only including participants who had passed two prerequisite undergraduate-level courses in civil engineering as well as a BK test ensured discipline-related BK. The results of the Reading module of IELTS were used for assigning participants into three proficiency groups. Participants took an EAP test consisting of four texts: two discipline-matching (civil engineering) and two discipline-mismatching (linguistics). Texts were designed to be of similar length and readability levels. Three task types were also constructed for each text classified as 'objective, semi-objective, and subjective' (Taghizadeh Vahed and Alavi 2020:6).

Results of the study indicated important roles for both language proficiency and discipline-related BK on EAP test performance. A facilitative role for BK was observed for intermediate and high-proficiency participants but not for the low-proficiency group, reflecting findings by Cai and Kunnan (2019). The interaction between discipline-related BK and task type indicated a non-significant effect of BK for objective tasks although a significant effect was found for the semi-objective and subjective task types. This is one of the only studies that emphasises the importance of taking task type into account when examining the role of BK in test performance.

Moving away from EAP/ESP domains, the positive role of culture-specific BK in enhancing comprehension, recall of information, and strategies used by readers has been documented in several studies (Carrell 1981, Johnson 1982, Malik 1990). 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 that subsume BK and interest in topic under one construct. To address this, the design of their research was such that the two variables would be separated in a study with 104 students of English as a Second Language (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 BK for the students so that each student could be administered texts in four conditions of high/low BK and high/low topic interest. To this end, 10 topics were selected from encyclopaedia entries, which ranged from computers, human evolution, and Islamic art to the Olympics and the stock exchange. BK 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 to 10. Proficiency level was not independently measured and instead, programme-level grouping was used as an indication of language ability. Reading comprehension was measured using 10 multiple-choice questions comprising both textually explicit and implicit questions. Once students had completed the BK 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 BK 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. Whereas the main effects for BK 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 BK and topic interest. It was shown that, if both are low, there is a significantly lower mean score for reading comprehension whereas higher BK and topic interest enhances reading comprehension to a small extent. The authors concluded that 'prior knowledge and topic interest are not highly significant nor additive factors in reading comprehension' (Carrell and Wise 1998:299).

Insights from listening comprehension research

The results of studies on the effects of BK 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' (1994:185). 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, 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; 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 acknowledged that the absence of an 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 BK in understanding lectures. In testing the hypotheses that listening proficiency moderates the effects of BK on lecture comprehension and that it is a strong predictor of lecture comprehension, data from 11 lecture subtests of a Test of Listening for Academic Purposes (T-LAP) was analysed for this study. Similar to one of Krekeler's (2006) measures, BK 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: BK and no BK. 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. As expected, a significant main effect was found for listening skills for all lectures. However, a significant main effect was found for BK 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 360 male-only Chinese English as a Foreign Language (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). 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 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' BK. Results indicated a significant interaction between BK 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 BK. 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).

The extent to which the study provides evidence in support of the facilitative effect of BK 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 baseline measure of BK and were once again administered after the lecture (with evidence of significant improvement in scores), then the authors would be better positioned in their interpretation. However, as it stands, the interpretation is largely speculative and the evidence from the study does not adequately lend support to a positive effect of BK on listening comprehension.

The use of topics that draw on markedly different levels of BK is a common method in examining the role of topic familiarity in experimental settings. The relevance of such an approach, however, can be questioned in LT contexts where differences in topics are far less pronounced. For example, the effects of religious-specific BK 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 BK 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 (the rock band U2). 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 BK 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' BK 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 to the positive influence of BK although 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 passage, leading 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).

A recent study by Li et al (2017) examined the effects of cultural familiarity and different question-previewing activities on the listening comprehension of 65 L2 junior high school students. A passage on the Moon Festival was considered culturally familiar whereas a passage on St Patrick's Day was considered unfamiliar. Efforts were taken to control for the difficulty levels of the passages. Participants were randomly assigned to one of two passages. Results suggested higher scores for the familiar text though differences failed to reach statistical significance, leading the authors to conclude that topic familiarity does not have an effect on listening comprehension. Note, however, that when participants were listening to the passages, texts were accompanied by a series of pictures. Moreover, prior to listening to the passages, key vocabulary items from the texts were also introduced. It can therefore be argued that these activities may have reduced (any) facilitative effect of familiarity and that perhaps too many confounding variables were introduced to the research making it difficult to draw substantive conclusions.

Insights from writing performance research

A clear trend for the positive role of BK on performance was found in Tedick (1990) where the written performance of 105 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; 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 BK. 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 these two prompts, Lim (2009:37) questions the 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 BK in examining the topic validity of an integrated test of reading, listening, and writing. In justifying this methodological approach, the authors point to the complexities of interactions between a given topic and test takers' prior knowledge of the topic, interest in the topic, and perceptions regarding the relevance of topics. They argue that test takers' choice of topic 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 two conditions: (a) no choice and (b) choice of five topics. Performances were subsequently compared across the two conditions on four dependent variables: 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 topic from the five available. However, due to this same design, there was no control over the number of participants selecting the said topic, which is why the number of participants in the two groups was not equivalent. Results of the Mann-Whitney U test showed 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. These non-significant findings are not unexpected, as the integrated nature of the test (see Chapter 1 for a brief discussion of independent and integrated tasks) designed to equip test takers with the prior knowledge necessary to complete the tasks potentially attenuates topic effects. These results were taken as validity evidence for the test in 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, test topics were perceived to be the second most important factor to influence 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 (Jennings et al 1999:449). 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 (similar to Swain's 1984 notion of 'bias for best').

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 that 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 BK of topics on performance scores was examined using multinomial logistic regression analysis. BK was not explicitly measured but was assumed on the basis of the students' departmental affiliation and candidates were subsequently divided into four categories: 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. This suggested that given a specific ability, the probability of achieving a given score would be dependent on the assigned topic, which would in turn 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 BK. I would also argue that that departmental affiliation is perhaps too broad a categorisation to be used as an indicator of BK or topic familiarity, as it is primarily assumption-based. The authors acknowledge that an independent measure of BK 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 potentially equip the test taker with the BK necessary to respond to prompts and thus reduce its effect on performance (Huang et al 2018, Jennings et al 1999, Weigle 2004).

In a study that focused primarily on general topics in a standardised assessment context, He and Shi (2012) raised 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' (2012:444). 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 demonstrated 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 and prompt type as well as a significant prompt–proficiency level interaction. The same trend was observed across the component scores. 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). Whereas 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' (He and Shi 2012:461) for further research on the subject.

Yang and Kim (2020) examined the effects of topic familiarity on features of CAF in the writing performance of 123 Chinese EFL college students. The two writing tasks used in the study are reproduced as follows (Yang and Kim 2020:87–88):

- "What do you think are the benefits and possible problems that computers and the Internet bring to university students in this country?"; more familiar (+)
- "What do you think are the benefits and possible problems that computers and the Internet bring to people in underdeveloped areas of the world where there is limited access to computers and the Internet?"; less familiar (−)

Familiarity in this research is operationalised as the extent to which there is a distance between the test taker and the subject matter, i.e. in relation to their country or in relation to a more unfamiliar group. The two tasks were randomly divided amongst the participants and a cloze test was used as a proficiency control measure. A post-task survey was also administered to check topic familiarity levels. Results of the study showed a significant effect of topic familiarity on CAF features collectively. Closer analysis revealed that it was only the lexical complexity measure that was influenced significantly with non-significant results for accuracy, fluency, and syntactic complexity. A possible explanation is that both tasks can be considered accessible despite the nuanced differences in their familiarity levels. The study could have also benefitted from a repeated measures design with the same participants performing the two tasks for more comparable results.

Insights from speaking performance research

Empirical research on the influence of topic and BK of topic in speaking performance contexts is rather limited and, similar to the other skills, 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' BK of the field-specific topics can be assumed, their level of BK of the general topics was not independently measured nor considered as an explanatory factor. Results of the study did not reveal a consistent trend: some ITAs performed 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). In examining the pass/fail cut-off scores, however, these small differences in scores could have made a difference between passing or failing for some of the ITAs in the study. The latter finding reflects Fulcher's (2003) notion of 'practical significance' discussed earlier in the chapter (see the section 'Insights from task-based language assessment research').

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, 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 stated 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 1992:323). This is one of the few studies which ascribes a negative influence to BK. It is also the first study discussed so far that briefly alludes to the potential role of rater characteristics in influencing performance assessment; a subject that I will discuss in more depth in Chapter 3.

Rather than distinguishing between field-specific vs. general topics, the focus of Papajohn's (1999) systematic and comprehensive study is the comparability of academic topics. Within the field of chemistry, the oral performances of 102 prospective ITAs on 15 different field-specific topics in the chemistry TEACH test were investigated. BK of topics was assumed for all chemistry graduate ITAs, given that the topics covered basic chemistry. Note 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, maths, 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 maths 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' (Papajohn 1999:62). 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. Two important factors may have confounded the study's findings: (a) teaching skills and (b) 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', which, in other words, is a case of construct conflation. 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, discussed earlier in the chapter, regarding 'difficulty' as a test artefact, rings particularly true in this example: where it is possible to question whether the difficulty of topics is an inherent feature of the tasks or whether, for example, it partly reflects raters' familiarity with some of the topics.

Inconsistent topic effects were reported in Lumley and O'Sullivan's (2005) study of a large-scale tape-mediated exit test of English in Hong Kong. The research examined the effects of task topic, test taker gender, and audience on the spoken performance of 894 university students. 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 Many-Facet Rasch Measurement (MFRM), the effects of each variable and their interaction on performance scores were examined. Scores were assigned using an analytic rating scale on the four criteria of task fulfilment and relevance, clarity of presentation, pronunciation, and grammar and vocabulary. 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 racing), there was a small advantage for males on the task fulfilment and relevance criterion with a female audience but this advantage was more pronounced with a male audience. 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). I would like to argue here that the classification of topics on the basis of such assumptions, that is according to preconceived notions of what is stereotypically male or female, may have confounded the results. A more critical issue perhaps is the group-level categorisation and analysis of gender as binary – male or female – which has long been a prevalent practice in the field of LT. While an in-depth discussion of gender is beyond the scope of this volume, it is important for the field to acknowledge and operationalise current conceptualisations of gender on a continuum, accounting for gender identities that transcend the traditional categorisations of male and female (Interagency Gender Working Group 2018[1]).

A strong positive influence for topical knowledge on spoken performance was found in Huang et al (2018). Comparing integrated and independent speaking tasks from TOEFL iBT preparation materials, Huang et al (2018) first developed and validated a series of topical knowledge tests, which were then administered to a group of 325 Taiwanese students of EFL. Participants were subsequently divided into two groups, with one group taking independent and integrated tasks on a specific topic combination and the other on a different topic combination. The results of path analyses suggested a statistically significant effect of topical knowledge on spoken performance across both independent and integrated task types and on the two topic combinations. This is one of the only studies that adopted

[1] www.igwg.org

a rigorous approach to measuring topic knowledge by constructing and extensively piloting a series of topical knowledge tests corresponding to the task topics used in the study. These tests were developed as a prediction measure; 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 BK. One possible limitation 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, although efforts were taken to minimise this by conducting data collection in two stages with a one-week interval. Nevertheless, the findings from this study are important not only in emphasising the role of topical knowledge but also demonstrating that the effects are not necessarily attenuated in integrated tasks, which counters previous claims in the literature (e.g. Read 1990, Weigle 2004). Huang et al (2018) explain this finding by referring to the possibility of the 'cumulative advantage' (DiPrete and Eirich 2006:271) of topical knowledge:

> The cumulative advantages from pre-task topical knowledge starts with an increased comprehension of the reading and listening input, which extends further to the speaking performance. Therefore, while the input offered by integrated tasks provides content for test-takers' oral responses, it might be more accessible to test-takers who comprehend input better due to pre-task topical knowledge. In other words, by supplying input, the integrated tasks might have somehow promoted the rich-get-richer phenomenon and allowed the presence of relevant topical knowledge to take on an even more facilitative/debilitative role in performance variations (Huang et al 2018:44).

Insights from IELTS-related research

In this section, I will be reviewing studies that have focused directly on the IST (see Chapter 1 for details of the test and its format). These largely fall into two categories: (a) research examining rater and test taker perceptions and experiences of the IST, and (b) the application of conversation analysis (CA) to IST performances. 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, studies directly focused on the IST and of relevance to the current research are considered first.

A concern with a potential topic effect in speaking tests was one of the strongest emerging themes in a worldwide survey of 269 IELTS examiners (Brown and Taylor 2006). This study was motivated by a major revision to the IST 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). Two of the questionnaire statements were particularly relevant: one related to examiners' views on the suitability of the topics for the candidature, and the other on the equivalence of topics in terms of difficulty. These statements were repeated for each of the three parts of the IST. Topics were found to be suitable by 55%, 70%, and 74% of the examiners for Parts 1, 2, and 3 of the test, respectively. Part 1 topics were found to be the 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 open comments touched on several issues regarding topics. Firstly, the incomparability of topics in terms of difficulty, appropriateness and/or complexity, and 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).

The examiners in the Brown and Taylor (2006) study also referred to topics in relation to the speaking assessment criteria. Emphasis is placed on these findings, as they relate to the rating scale that will be used as an instrument in the main study (see Chapter 4) with important methodological implications. 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 2006:16–17). 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 and at the time of writing, these recommendations have not been implemented in the scales. Secondly, in respect of 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 also strongly resonates with the results of Brown's (2006) qualitative enquiry into the validity of the IELTS rating scales discussed further below. 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. 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).

Brown (2006) analysed the verbal reports of six IELTS examiners as they rated previously recorded speaking tests using stimulated recall methodology 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 the questionnaire (Brown 2006:8–9).

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 IST 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 a higher education institute in the UK. 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.

Topic, once again, was found to be a recurring issue in the 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 (seven 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 Participant I4 in Smith's study (2009:59) 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.

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 discussed earlier in the chapter.

Inoue et al (2021) conducted a large-scale survey with 1,203 IELTS speaking examiners and examiner trainers in order to gather their views and voices regarding the current format of the test and to explore possible suggestions for future improvements to the test. When asked if topics in the test tasks are appropriate, approximately 40% of respondents disagreed or felt neutral about the overall appropriateness of the topics. When narrowed further, results suggested that topic appropriateness was problematic in relation to candidates' cultural background and gender. To follow up on survey results, semi-structured interviews were conducted with a small sample of examiners (n = 36) in order to get a more in-depth understanding of the views expressed.

A recurrent theme was the 'incongruity of a given topic within a specific cultural context'; for example, music or pop stars were found to be inappropriate in some countries in the Gulf or the Middle East or specific modes of transport such as bicycles or boats were found to be problematic in contexts such as Saudi Arabia or central China. Examiners also commented on the affective nature of some topics with some candidates 'breaking down' when asked to talk about family members or a past memory. Issues of class and socio-economic status were raised by several examiners who referred to some topics as too 'middle class' or outside the experience of candidates from lower socio-economic backgrounds (Inoue et al 2021). Some topics and questions were also found to be controversial by examiners in 'reinforcing gender stereotypes' although others found them unproblematic given the more 'traditional' settings they were examining in. Linked to these issues were the examiner script and standardisation requirements, which were viewed as not allowing examiners the flexibility to 'intervene' or 'take appropriate actions when the test does not proceed smoothly and as intended'. These results strongly echo the findings from Brown and Taylor's (2006) survey discussed earlier, suggesting fundamental topic-related issues with the IST that have persisted for nearly two decades and since the last major revision to the IST in 2001.

The studies examined so far have focused on stakeholder perceptions of IST and issues related to topics from the *outside*. Let us turn attention to research that has examined the role of topic from a structural perspective in organising interaction *within* the test.

CA has been used to study topic within the interactional organisation of the IST focusing on how 'topic initialisations, shifts, and endings are managed as an interactional achievement in the unfolding of the moment-to-moment interaction from the participants' perspective' (Seedhouse 2018:115). Management of topic in CA research more broadly has been categorised into distinct 'stepwise' (Sacks 1992:566) or 'marked' (Sacks 1992:352) organisations, with the former suggesting a 'flow' from one topic to another whereas the latter involves more explicit shifting of topics and a 'larger distance between the topics than in stepwise transitions' (Seedhouse 2018:115). These distinctions are particularly relevant to the IST: the design of the test and its format combined with the strict examiner frame can result in more 'marked' or 'boundaried topic shifts' (Seedhouse and Harris 2011:8) which may detract from the quality of good conversation (Sacks 1992:352).

In a series of studies by Seedhouse and colleagues (Seedhouse and Egbert 2006, Seedhouse and Harris 2011, Seedhouse 2018) CA was applied to transcripts of IST performances in order to explore the interactional organisation of the test and the ways in which topic is developed throughout the test. Using an institutional discourse perspective, the methodology 'attempts to understand the organisation of the interaction as being rationally derived from the core institutional goal' (Seedhouse and Harris 2011:6). Here I discuss some of the main findings from these studies relevant to the focus of this volume.

Seedhouse and Egbert (2006) applied CA to 137 IST transcripts and found the overall organisation and turn-taking of interaction in the IST to closely align with examiner instructions and the 'institutional aim of standardisation' as the key organising principle of the interaction (Seedhouse and Egbert 2006:1). Examiners nominated topics in line with the script; there were very few occasions of candidates attempting to nominate topics and those were nevertheless denied by the examiner (Seedhouse and Egbert 2006). These findings were taken to suggest that the organisation of the IST differed significantly from ordinary conversation and closely resembled 'goal-oriented institutional interaction' (Seedhouse and Egbert 2006:32). Their data also revealed problematic questions or sequences of questions on a specific topic particularly in cases where there was 'an unmotivated and unprepared shift in perspective of any kind' (Seedhouse and Egbert 2006:35) with a recommendation to pilot topics and question sequences.

Seedhouse and Harris' (2011) CA study explored the ways in which topic is developed in the IST using a corpus of 60 transcribed IST performances. Findings from the study revealed topic as an essential aspect of the IST and 'inextricably entwined with the organisation of turn-taking, sequence and repair and as directly related to the institutional goal' (Seedhouse and Harris 2011:37). Examiners were shown to use different interactional resources to mark shifts in topic with variations according to different test parts. In

Part 1, for example, topic boundary markers (TBMs) were predominantly determined by the examiner script whereas in Parts 2 and 3 of the test examiners employed a range of TBMs categorised into 'unmarked, generic, and explicit' (Seedhouse and Harris 2011:34). Importantly, the data demonstrated how the management of topic is exclusively determined by the examiner and the examiner script with 'asymmetrical rights to topic management between the examiner and the candidate' (Seedhouse and Harris 2011:37). Development of interaction was shown to follow an archetype with all topics introduced by means of a question posed by the examiner with questions containing 'an adjacency pair component' requiring a response from the candidate and 'a topic component' requiring the development of a topic by the candidate (Seedhouse and Harris 2011:38). These features were taken by the authors to align with the institutional goal of ensuring standardisation and validity of assessment (Seedhouse and Harris 2011). Reference, however, was also made to problematic sequences in the data where specific questions were shown to be challenging even for high-scoring test takers. These were found to be largely due to questions that 'may involve an unmotivated shift in perspective, may require special knowledge or experience which may not be available to most candidates, or may be puzzling in some way' (Seedhouse and Harris 2011:38). These findings substantiate some of the concerns raised by the examiners in Brown and Taylor (2006) and Inoue et al (2021).

Seedhouse (2018) builds on the work of Schegloff (2007) and Heritage (2012) and the data from Seedhouse and Harris (2011) to illustrate how topic has developed a 'dual personality' (2018:114) in the IST. Schegloff's (2007:1) research suggests treating topic in relation to 'action' rather than 'topicality' and Seedhouse (2018) therefore analyses the ways in which topic might facilitate institutional action. The work of Heritage (2012) suggests a consideration of the 'epistemic engine' of talk where imbalances in information 'drive' talk until balance is reached (cited in Seedhouse 2018:116). This is applied by Seedhouse (2018) to the IST given the information-transfer function of tasks used in the test. In his analysis, Seedhouse (2018:116) views the two functions of topic in serving the institutional goal as follows:

> Topic-as-script is the homogenised topic which examiners give to candidates, whereas topic-as-action refers to the diverse ways in which candidates talk a topic into being. The movement from "topic" as a single homogeneous script to a heterogeneous series of responses by different candidates enables differential ratings of their performances.

Topic therefore is viewed as one of the most important aspects of the IST in driving the interaction, ensuring standardisation, and allowing the elicitation of different levels of performance for scoring. A critical issue that remains unaddressed in these studies is whether such alignment with the institutional goal of standardisation is at the expense of construct validity when elicited interactions fail to exhibit aspects of more natural and authentic interactions in real life.

Insights for future research design

So far in this chapter I have provided a critical review of research that has focused on the effects of topic and BK of topic on performance. Findings from different studies have been shown to be inconsistent and largely inconclusive in relation to topic/BK effects. Why might this be the case? Below are some thoughts and possible explanations.

Operationalisation of BK

One possible explanation is the different ways in which BK has been operationalised; for example, it has either been *assumed* on the basis of factors such as current or future academic field of study (Alderson and Urquhart 1983, 1985, Lee and Anderson 2007), cultural background (Chiang and Dunkel 1992, Li et al 2017), and gender (Lumley and O'Sullivan 2005) or *inferred* on the basis of a variety of methods such as self-report questionnaires (Carrell and Wise 1998), previous experience with topics/texts (Schmidt-Rinehart 1994), reading habits (Clapham 1996), or topic knowledge tests (Huang et al 2018, Usó-Juan 2006).

What the literature has revealed is a superiority of inference-based methods of establishing BK over assumption-based ones. This is given the decidedly individual nature of BK, which does not lend itself easily to assumptions, generalisations, and stereotyping.

Proficiency level as a key factor

Another important insight from the review of the literature is the differential way in which BK may affect performances of test takers from various levels of proficiency (Cai and Kunnan 2019, Krekeler 2006). Proficiency level, however, similar to BK, has been operationalised and measured in various ways, which can explain why results of studies are inconsistent and, at times, contradictory. Amongst the different measures used are self-reports of proficiency (Long 1990), course-level information (Carrell and Wise 1998, Schmidt-Rinehart 1994), grammar tests (Cai and Kunnan 2019), C-tests (Krekeler 2006), and standardised tests such as TOEFL (Lee and Anderson 2007).

Influence of other variables

In addition to proficiency level as an important variable, the review of the literature has pointed to other factors which may need to be taken into account such as task type (Taghizadeh Vahed and Alavi 2020) or raters (Douglas and Selinker 1992, Papajohn 1999) as well as the importance of adopting methods that allow for the systematic separation of parameters of interest (Bachman 2002) such as topic (as a task parameter) from BK of topic (a test taker parameter).

Implications

Drawing on the review of the literature in this chapter, a number of important implications have emerged for the main research: firstly, a non-assumption-based measure of BK needs to be included in the research. Secondly, proficiency level is a key variable to be incorporated in the research and measured with a reliable instrument. I will address both points in more depth in Chapter 4. Lastly, a measurement model is required that allows for different parameters of interest (topic and BK of topic) to be conceptualised independently while taking into account the influence of other variables such as raters or task types. This is the subject of the next chapter.

3 Networks of interaction: Measuring and judging performance

In the previous chapter we looked at the literature on the effects of two related factors on performance in an L2: topic as a task characteristic and BK of topic as a test taker characteristic. In this chapter, I will focus on the role of a third key factor – *raters* – briefly referred to in the previous chapter. The influence of raters is inextricably linked to the assessment process and is a source of variability, the importance of which cannot be understated. I will then turn to a consideration of how different aspects of a testing situation related to tasks, test takers, and raters are at play in contributing to a score observation and how we can go about measuring and reporting performance.

Raters: The 'Achilles heel'[1] of performance assessment?

Performance assessment 'necessarily involves subjective judgements. This is appropriate: evaluation of any complex human performance can hardly be done automatically' (McNamara 1996:177)[2]. The pivotal role of raters lies in the link they create between a test taker's performance, rating scale(s), and scores. By involving human judges however, a degree of subjectivity or an 'Achilles heel' (O'Sullivan and Rignall 2007:447) is simultaneously introduced to the assessment process. Issues subsequently arise from the interdependence of such subjectivity with measurement error (Cronbach 1990), most critically associated with relative rater harshness and leniency,

[1] The term 'Achilles heel' refers to a weakness or vulnerability and is rooted in Greek mythology. O'Sullivan and Rignall (2007:447) used the term in relation to the (undesirable) variability in performance assessment that can be introduced by subjective human judgements.

[2] Since McNamara's statement in the late 1990s, there have been various advances in speech processing and machine learning technologies that have allowed for the development of automated speaking assessment where candidate responses are scored by computer algorithms rather than trained human raters (Chapelle and Chung 2010, Wang, Zechner and Sun 2018). Examples include the Versant test and TOEFL Go. Although a discussion of these tests and the automated approach to scoring is outside the scope of this volume, a point worth emphasising is that even the most sophisticated of these automated systems have limited capacity in capturing high-level features of speech such as content appropriateness, topic development, and discourse organisation (Chen et al 2018). More than two decades on, McNamara's point perhaps still stands.

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 (Bachman et al 1995, Eckes 2019, McNamara 1996, Wiseman 2012). The quote below from around the turn of the 20th century establishes the rater effect as a longstanding phenomenon:

> 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 (Edgeworth 1890:653).

Notice how the element of 'chance', that is, the choice of judges who 'happened to be appointed' is carefully distinguished from the 'competence' of the judges. It is not the judges' ability or expertise that is in question. Rather, it is their individual differences, likely in terms of harshness and leniency, that are considered to contribute to large variations in scores or affect the probability of success or failure. This observation, from well over a century ago, reflects findings from recent studies, which largely suggest that the rater effect is here to stay.

Is rater training the answer?

The often-cited solution to reducing the rater effect is through training, and yet, the literature on rater training for both writing and speaking assessment has consistently shown that whilst intra-rater reliability can be improved with training, the influence on improving inter-rater reliability is far from ideal. Weigle's (1998) study, for example, systematically analysed score data from 60 written essays using MFRM, with the overall aim of distinguishing between the effects of rater training on rating consistency and relative harshness or 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. Training was viewed positively in 'helping raters give more predictable scores' but *not* in 'getting them to give identical scores' (Weigle 1998:263). In other words, training was found to be effective on improving intra-rater reliability as opposed to inter-rater reliability. A similar conclusion was reached in Lumley and McNamara's (1995) longitudinal study of rater training in the context of the speaking subtest of the Occupational English Test (OET). Ratings were 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' (Lumley and McNamara 1995:69).

Whereas Lumley and McNamara's (1995) study was limited to experienced raters in a speaking context, Lim's (2011) large-scale study included both novice and experienced raters in writing assessment. The research focused on raters' scoring behaviour, 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).

An alternative approach to decreasing examiner variation was investigated in O'Sullivan and Rignall (2007). The study in the context of the IELTS Writing module explored the usefulness of providing formal feedback (from the results of MFRM bias analyses) to trained IELTS raters in order to improve 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. Questionnaire results from raters, on the other hand, revealed a positive and motivating impact of the feedback on raters' decision-making processes. Drawing on these findings, where even experienced and trained examiners exhibited instances of bias, the authors questioned 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).

Other studies using MFRM analyses have also revealed large severity differences amongst IELTS examiners. Brown and Hill (1998), for example, reported a severity difference of 0.6 of a band for the six speaking examiners in their study. Note that this was for the pre-2001 version of IELTS. Nakatsuhara, Inoue, Berry and Galaczi (2017) and Berry, Nakatsuhara, Inoue and Galaczi (2018) reported severity differences of 0.52 and 0.76 of a band for the 10 and eight examiners in their speaking studies, respectively. In the context of IELTS, half a band can have practical significance for test takers, echoing Edgeworth's concern with consequences.

Taken together, this body of research points to significant rater effects on performance while highlighting the limited effectiveness of training on reducing rater differences in both speaking and writing assessment. Let us now consider three important issues in light of the consistency and stability of the above findings. The first is to question the desirability of 'perfect agreement' (McNamara 1996:126), which turns the problem of rater variation on its head. The second is a consideration of *why* training efforts have been unsuccessful in dramatically reducing rater variability in assessment contexts, and the third is how to deal with such variation.

The allure of perfect agreement

Perfect rater agreement has been problematised by Constable and Andrich (1984, cited in Lumley and McNamara 1995:56) who cautioned that such reliability might only 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 that is 'true'. The term refers to the subjectivity of perception and borrows its name from a film title by Akira Kurosawa, a master of classic Japanese cinema. In *Rashomon*, a crime is witnessed by four observers who go on to describe the event in four 'contradictory yet equivocal' ways (Fanselow 1977:17). The term has since come to 'embody a general cultural notion of the relativity of truth' (Kamir 2000:41). In relating it to performance 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. Similarly, Lim (2011:557) asks 'who is to say whose judgments are better or worse? And does that not require judgment as well?'. The answer, according to McNamara, is to accept rater variation as 'a fact of life', direct training efforts towards increasing raters' internal consistency, and to compensate for such differences through alternative approaches such as multiple marking or more sophisticated measurement models such as MFRM. In MFRM, raters are conceptualised and subsequently modelled not as 'scoring machines' but as 'independent experts' (Linacre 2018a) where the individual differences between raters are not viewed as hindrances to be eliminated but rather as expected features of the rating process that can be modelled and subsequently accounted for (Weigle 1998).

The intricacies of the rating process

An important issue to explore is the reason why training appears to be less effective than desired in reducing rater subjectivity. One explanation lies in the complexity and cognitive demand of the rating process (Cronbach 1949, Wiseman 2012) and the different factors that may affect decision making. These may include rater characteristics such as their L1 (Kim 2009), rating experience (Lim 2011), academic discipline (Vann, Lorenz and Meyer 1984), or the extent of rater engagement or 'ego-involvement' (Myford and Wolfe 2003, Wiseman 2012) where through a process of identifying with or 'personalising' (Wiseman 2012:151) a candidate or a performance, the rater may display a distinct scoring behaviour. Raters may also differ in how they interpret or understand a scale or the features of performance they find particularly salient (Eckes 2019, Pollitt and Murray 1996). These multiple rater tendencies may manifest themselves in a number of ways – the most prominent of which are identified below (Linacre 2018a, Myford and Wolfe 2003, 2004):

- 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 more similar ratings than justified on different criteria of an analytic rating scale (Marais and Andrich 2011)
- the extreme/central tendency effect, which refers to a rater's tendency to award scores in either the extremes or middle categories of a rating scale (Wolfe, Jiao and Song 2015)
- the 'playing it safe' effect (Linacre 2018a), which refers to a rater's disposition towards assigning scores which are similar to those awarded by another rater
- the 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 2019).

Distinct from systematic rater effects is the 'random effect', which refers to unpredictable or 'haphazard rating patterns' (Wind 2019:4) that cannot be explained in a systematic fashion (McNamara 1996). This is also referred to in the literature as 'rater inaccuracy' (Wolfe and McVay 2012) or 'noisy ratings' (Wind and Engelhard 2013).

This discussion has shed some light on why training may not always be successful in reducing rater variation, as it would be challenging, if not impossible, to individually identify, isolate, and address this multitude of tendencies during rater training. How then might we address the problem?

Measurement of performance

In McNamara's (1996) model of test performance (see Chapter 1), various aspects of a testing situation, also known as *facets*, were shown to be at play in contributing to a score observation (see also Milanovic and Saville (Eds) 1996 for a comprehensive overview of facets involved in large-scale performance testing). Raters, test takers, tasks, and their different characteristics are all examples of facets. A central issue is how to isolate the relative influence of these facets from one another. In other words, how can one determine whether a given score reflects the 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?

One possible solution is the use of MFRM (Linacre 1989). The strength of MFRM is that it allows for various facets of the testing event to be modelled simultaneously but examined independently. As such, it addresses various psychometric concerns that have been raised in relation to the measurement of performance within the traditions of classical test theory where sample-dependent approaches to measurement and reliance on raw scores as an indication of test takers' underlying ability have been found to be wanting. Bachman (2004:152) summarises some of these 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.

The three most common IRT models in modern test theory as listed by Bachman (2004:152) are: 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 this volume, I will be using the Rasch model and its extension to MFRM (Linacre 1989), which is now a widely practised approach in rater-mediated contexts that involve constructed-response items and the awarding of scores using rating scales (Eckes 2019, McNamara et al 2019). Before explaining the model in more detail, I would like to draw attention to a significant source of controversy between those who see the Rasch model as a special case of IRT classes of models (e.g. Bachman 2004) and those who see the Rasch model as substantially different[3]. 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.

Approaches to measurement: A case of different paradigms?

A *paradigm* (Kuhn 1962) 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 the words of Andrich (2004:I-7), it 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 enquiry – 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, those who subscribe to the 'traditional paradigm' (Andrich 2004:I-7) view the Rasch model as a special case of IRT classes of models. In this paradigm, 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: should the one-parameter model not fit the data, then the two-parameter model can be used and so on and so forth. In contrast, in the 'Rasch paradigm' (Andrich 2004:I-7), there is an *a priori* constraint on the class of models and parameters that can be used, that is, only those models that 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 needs to fit the model (Andrich 2004). When there is evidence of data misfit, the Rasch paradigm invites a qualitative enquiry into the original data, rather than opting for other increasingly sophisticated models. In this volume, the approach to

[3] See Andrich (2004) for a comprehensive account of the origins of the controversy.

research and the data–model relationship will be conforming to the Rasch paradigm. I will now describe the Rasch model and its extension to MFRM in more detail and illustrate how the model can be applied to the areas of enquiry in this volume.

The Rasch model: An elegant solution?

Rasch analysis is the 'formal testing of an outcome scale against a mathematical measurement model' and was developed by the Danish mathematician Georg Rasch (Tennant and Conaghan 2007:1,358). 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)[4] which holds important implications for the social sciences: '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 is rarely deterministic. Conversely, the Rasch model uses a probabilistic counterpart of Guttman scaling[5], which meets the criteria of fundamental measurement, to provide a framework against which data from the social sciences can be

[4] '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; emphases in original).

[5] Guttman scaling is a 'deterministic pattern that expects a strict hierarchical ordering of items' (Tennant and Conaghan 2007:1,358). 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).

formally tested (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.

Note that one of the main underlying assumptions[6] 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 classical test theory (Andrich 1982:95), where items in a test are designed to reflect the same thing (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, raising this question might be misleading in itself: '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 difficulty of an item (δ_i). If β_v is larger, equal to or smaller than δ_i, then we can expect the following probabilities[7]:

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

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 the following form:

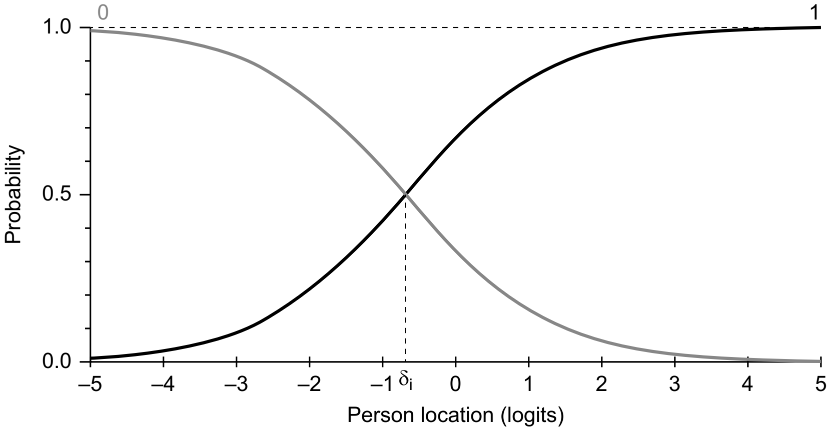
$$P\{x_{vi} = 1 | \beta_v, \delta_i\} = \frac{e^{(\beta_v - \delta_i)}}{1 + e^{(\beta_v - \delta_i)}}$$

[6] The other main assumption is 'response independence' which I will cover later in the volume.

[7] The formulae in these sections are reproduced from Andrich and Marais (2010).

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 3.1.

Figure 3.1 Item characteristic curves for a dichotomously scored item (*i*)



The horizontal line represents *person ability* in logits[8] whereas the vertical line displays *probability values*. The light grey curve represents the model's expectations for a score of 0 while the black curve represents expectations for a score of 1. As we move along the person ability continuum, or as ability increases, the probability of getting a score of 0 decreases while the probability of getting a correct answer of 1 increases. The location at which the difficulty of the item is equal to the ability of the person (marked with a 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, that is items with more than two response categories. You will see applications of these models to various instruments later in this volume.

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

Another extension of the Rasch model is MFRM (Linacre 1989), and is expressed as follows:

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

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 MFRM, the person parameter can be conditioned out and parameters for rater and item (or other facets such as prompt, task, criterion) can be estimated simultaneously. 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 one another (Eckes 2009). Examinee scores can be adjusted for the impact of other facets accordingly, thus presenting a more accurate measure of examinee ability and providing an elegant solution for measuring performance.

In the next chapter, I will illustrate how the Rasch family of models was used as the main method of analysis for both evaluating the quality and functioning of different research instruments and for examining the influence of the main facets of interest – topic and BK of topic – independent of but adjusted for the influence of other facets.

4 Investigating topic and background knowledge of topic: A research study

In this chapter, I will describe and explain in detail the research design, methodology, instruments, and procedures that were used to gather evidence for the topic validity of an L2 speaking test. To remind the reader, the focus of this volume is on the extent to which topic and BK of topic have an effect on spoken performance. I have argued that in assessment contexts where topics are randomly assigned to test takers, it is important to demonstrate that topics of tasks and the level of BK that test takers bring to these topics do not exert an undue influence on test results. Otherwise, a validity threat may be introduced to the test. The following RQ guides the study: *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?*

To facilitate the systematic collection of different types of validity evidence (in line with Weir 2005), this overarching RQ is broken down into a series of specific subsidiary RQs:

i. *To what extent are the topics of 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 differences in test takers' levels of BK of topics have an impact on performance? Are (any) differences large enough to have practical significance in terms of test performance?*

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

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

ix. *What themes and patterns emerge from an enquiry into (a) rater perspectives (b) the content of test taker speaking performances, and (c) test taker perspectives in relation to topic validity aspects of the speaking test under examination?*

A mixed methods approach

In this study, I employed a mixed methods approach to research where both quantitative and qualitative methods and strategies were brought together as appropriate. These strategies of enquiry align 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 research findings. However, in an embedded design, the two strands of enquiry 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). In examining the effects of topic and BK of topic on performance, the validity argument of my research rests predominantly on quantitative data with qualitative information serving to enrich the findings and providing a deeper understanding of the research problem (Creswell and Creswell 2017, Tashakkori and Teddlie 2010).

The majority of subsidiary RQs in my study are addressed using quantitative techniques, involving the analysis of scores and closed-ended questionnaire responses. The study's final subsidiary RQ, however, involved qualitative analysis of the content of test taker performances and raters' interview transcripts. The data analyses for the two strands of enquiry were carried out independently but were subsequently 'mixed' and synthesised in the interpretation phase in order to address the overarching RQ.

Research setting

The EFL context of Iran serves as the backdrop for this research for two reasons: firstly, my own familiarity with the context, which helped facilitate data collection. Secondly, because of the important role of IELTS in Iran for allowing access to academic and immigration opportunities. This is evidenced in the IELTS global test statistics showing Iran to be consistently amongst the most frequent test-taking nationalities[1]. In addition, empirical research from local assessment settings such as Iran contributes evidence in relation to the 'local' validity of IELTS, that is the extent to which a global test is appropriate (or not) for a given local context. 'Localisation' places emphasis on the test taker in the context of the social domain and explicitly recognises test consequence as a central consideration in test development and validation (O'Sullivan 2016).

Participants

There were two groups of participants in this study: test takers and raters.

Test takers

82 adult non-native speakers of English (L1 Persian), aged between 18 and 40, agreed to participate in my research. Half the participants identified as female and half as male. All were enrolled in English language courses and were preparing to take the IELTS exam. They were familiar with the general format of the IELTS Speaking test (IST) as confirmed by their teachers. This was a fairly homogenous sample in terms of L1, cultural background, and exposure to the target culture.

To select participants, I contacted language schools and private institutions that offered IELTS test preparation courses and invited them to collaborate in the research project via email. I enclosed general information about the study such as its purpose, what participation would entail, and the data collection schedule as well as documentation regarding ethical approval by Oxford University's Central University Research Ethics Committee.

Once access was granted by the institutes, I informed students enrolled in IELTS preparation courses about my research and invited them to participate in the study. I did not disclose the exact nature of the study in order to allow for spontaneous speaking test performance. They were informed that there would be an opportunity to practise mock speaking tests with a former IELTS examiner and that I would provide them with feedback on 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. There was a lot of enthusiasm to participate in the study, as students saw this as an opportunity for extra speaking practice. Participation was strictly on a voluntary basis and no honorariums were provided. Interested students signed their consent forms and we agreed on a schedule for data collection

[1] www.ielts.org/for-researchers/test-statistics/test-taker-performance

in consultation with teachers and the language school administration office.

Raters

Four raters participated in this study. Three identified as female and one as male. All reported English as their L1. I based rater selection on the following criteria: (a) familiarity with language proficiency interviews such as IELTS and TOEFL, (b) English language teaching and/or examining experience (more than five years), and (c) a graduate degree in Teaching English to Speakers of Other Languages (TESOL), applied linguistics, language assessment or a related field of study. All raters were provided with training prior to the rating process.

Parallel forms design

Before selecting speaking tasks for the research, it was important to establish the number of tasks/topics that could be feasibly included: 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 introducing fatigue or boredom.

I decided to opt for a parallel forms reliability design where each participant responds to two parallel/alternative versions of the IST, each consisting of five topics. To remind the reader, each IST is comprised of three task types (identified as 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 Chapter 1 for detailed information about the IST and its format). 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 tests, that is a complete or fully crossed design (Eckes 2009), I chose a more practical solution in an incomplete design (Eckes 2009, Weir and Wu 2006). An incomplete design necessitates the linking of the two tests through *common topics* but simultaneously allows for an increase in the number of topics that can be included. This design-related decision was possible in light of the measurement model selected for the study – 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 research, I developed four alternate versions of the IST with two common tasks creating the necessary common link between the tests, allowing for coverage of 18 different topics. Table 4.1 illustrates the incomplete-connected data collection design of the study. Topics A.1 and A.2 (in bold) are the common items.

Table 4.1 Incomplete-connected research design

Topics	Task types	Test takers (Group 1)	Test takers (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

Speaking tasks

Speaking tasks were selected from a pool of publicly available IELTS materials. Task selection followed two main steps: firstly, a review of published ‘retired’ IELTS papers that included authentic tasks from previous IELTS administrations. 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 such as IELTS and TOEFL. Retired IELTS materials previously used in live operational settings served this purpose. Following a preliminary review, a total of eight Task Type A topics, five Task Type B topics, and 10 Task Type C topics (thematically linked to Task Type B topics) were selected.

The second step consisted of an in-depth analysis of tasks in order to establish task equivalence. In line with Bachman (2002) and Weir, O’Sullivan and Horai (2006), it was important to ascertain that with the exception of task topic, other task-related variables were controlled for as far as possible so that (any) differences in scores could be predominantly attributed to differences in task topics and test takers' BK of topics.

In the review of the literature, several task processing conditions such as code complexity, cognitive complexity, and communicative stress (Skehan and Foster 1997) were identified as factors that can affect L2 performance. Weir et al (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 this study and the condition that varies across tasks, the remaining conditions had to be addressed and controlled for.

Planning time, planning condition, and response time are factors that can be controlled through test administration procedures and by adhering to a strict interview structure. For example, similar to the live test, providing a specific timeframe 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. The potential effects of audience and interlocutor-related factors can be controlled by having the same examiner/interlocutor, and by closely adhering to the IST format, the type of input in different parts of the test can be held constant for all participants.

Amount of input was examined both quantitatively and qualitatively in order to ensure task equivalence in relation to level of input. The quantitative analysis of tasks involved the calculation of descriptive statistics in terms of average number of questions per task, words per sentence, and characters per word. For Task Type B options (written prompts), the Flesch Reading Ease and Flesch-Kincaid Grade Level were also calculated and taken into account in the selection process, following Weir et al (2006). The qualitative examination of the tasks was carried out by inviting a select panel of experts with specialisation in applied linguistics and second language acquisition (SLA) to rate the tasks. The panel was asked to apply a task equivalence checklist adapted from Weir et al (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 (a copy of the checklist is provided in Appendix A).

The final task selection decision was made on the basis of a number of different types of evidence: the panel's ratings of tasks, the qualitative analysis of task-related comments, and a consideration of task input statistics. Tasks flagged for containing relatively unfamiliar grammar or lexis were removed. On the other hand, tasks that were considered to be comparatively 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 that exhibited equivalence in terms of features to an acceptably high degree. The final set of speaking tasks chosen for the research can be accessed in Appendix B. These tasks were divided into four alternate speaking test forms – W, X, Y, and Z – according to the IST format (see Appendix C).

A measure for background knowledge

In the review of the literature in Chapter 2, I outlined the various methods with which the construct of BK or topic familiarity has been operationalised in research in the field. One approach has been to use extreme levels of familiarity; for example, by matching/mismatching individuals to specific culture-laden topics or academic backgrounds. Another approach involves establishing familiarity using binary yes or no questions or using topic knowledge tests. In some studies, BK was simply assumed or inferred without an independent measure. As I argued in Chapter 2, however, the interaction between a test taker's BK and topic of a task is a complex phenomenon which cannot be assumed or pre-determined. The solution I opted for was to include an independent measure of BK – a questionnaire – that allows for capturing of different levels of topic-related BK while taking into account the possible influence of other topic-related factors such as interest in topics or perceptions of topic difficulty. The parallel forms design of the study, with each participant responding to a range of different topics, would also allow participants to gauge their relative degree of BK related to each topic.

A post-speaking test questionnaire was therefore constructed to elicit test takers' self-reports of BK of topics assigned to them. The BK questionnaire consisted of three main sections. I will describe each section in more detail and provide a rationale for design choices.

Section I. This section consists of eight questions (see Table 4.2) repeated for each topic. Responses were elicited on a five-point Likert scale from strongly disagree (1) to strongly agree (5) with an undecided option (3). A space was provided for open-ended comments.

Table 4.2 Section I questions

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.

Questions were formulated on the basis of the literature and following a review of available instruments (e.g. Weir 2005, Weir et al 2006). Questions 2, 3, 4, and 6, for example, were adapted from Weir's (2005:237–239) cognitive processing questionnaire. It was hypothesised that a higher degree of familiarity/BK would correspond with an easier production of ideas from memory (Question 4).

Some questions are more directly associated with BK and were designed to elicit participants' self-reports of degree of topic familiarity and ideas about a topic. Others tap into topic-related factors such as interest in topic and perceived topic difficulty in line with the literature (Carrell and Wise 1998, Jennings et al 1999). Question 5 attempts to tease apart language ability and BK.

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 with eight questions each).

Section II. This section of the questionnaire focused on test taker perceptions of the role of topic and topic familiarity on their performance scores and the importance they place on these factors. The four questions in this section are reproduced below as Table 4.3:

Table 4.3 Section II questions

1. I think that the choice of topics might affect my final score.
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 topics is not important if my English is good enough.

Whereas Questions 1 and 2 focus attention on the effects of topic and BK 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' BK 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 elicited some general information from participants such as age, gender, education levels, self-reports of English proficiency, and familiarity with the IST format. This section was placed last, as it poses less cognitive demand on the participants, requiring only personal factual information.

A measure for language proficiency

The review of the literature highlighted the role of general language proficiency as an important variable that can shape the way in which test takers' BK of topics interacts with task topics. To examine this potential interaction in my research, I opted for C-tests as a measure of general language proficiency. The choice of instrument followed a consideration of different measures of proficiency as well as practical constraints. Importantly, the literature on C-tests provided strong support for their construct validity; for a comprehensive review see Eckes and Grotjahn (2006). C-tests are also easy to develop, administer, and score thus making them an appealing option as a research instrument and control measure for general language proficiency. In the next sections, I will provide a brief introduction to C-tests and describe the steps taken for developing, piloting, and validating a set of C-tests for this study.

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).

The construct validity of C-tests

What C-tests measure and their construct validity has been the subject of much debate. On the one hand are those who view C-tests as having limited functionality in assessing micro-level processing only (Cohen, Segal and Bar-Siman-To 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 knowledge of grammar and vocabulary is predominantly targeted by these types of tests (Kamimoto 1992). On the other hand, there is contrary empirical evidence that shows performance on the C-tests to be 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 a systematic study examining the construct validity of C-tests, Eckes and Grotjahn (2006) employed two statistical analytical approaches – Rasch analysis and confirmatory factor analysis – to determine whether the trait underlying performances on C-tests was the same as that underlying performance on a standardised large-scale test of German (Test Deutsch als Fremdsprache, known as TestDaF) and its subsections. Their findings provided 'clear evidence' that the C-test measured the same general dimension as TestDaF (Eckes and Grotjahn 2006:290). Singleton and Singleton (2002) also showed high correlations between scores on C-tests and the productive skill of speaking. Drawing on the body of research on C-tests, Hastings (2002) asserts 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' (cited in Eckes and Grotjahn 2006:292).

C-test design and piloting

In designing the study's C-test, several resources (Klein-Braley 1997, Klein-Braley and Raatz 1984) were consulted for guidelines and instructions on C-test construction and the following steps were taken:

- text selection – a set of 15 texts were selected initially
- establishing text difficulty using the Lexile-measure[2]
- text conversion to C-test format
- C-test piloting (with all texts) with proficient users of English
- modification/elimination of texts following pilot study results
- piloting with users of English from a range of proficiency levels
- selecting texts for the final C-test
- developing an answer key on the basis of the pilot study responses.

The final C-tests (three versions) are provided in Appendix D. A common-item linking approach was used in order to include a larger selection of texts and to increase test reliability.

Rating scales

Luoma (2004) invites researchers to use existing rating scales and to tailor them according to their own research needs. For this study, I used the public version of the IELTS Speaking Band Descriptors[3] to score spoken performances. This analytic rating scale has been extensively used

[2] The Lexile measure (or L-measure) is a Rasch-calibrated measure that gives objective information about the difficulty of a text or a person's reading ability. It is calculated on the basis of two strong predictors of text difficulty: word frequency and sentence length (metametricsinc.com/parents-and-students/lexile-for-parents-and-students/lexile-for-reading/).

[3] www.ielts.org/-/media/pdfs/speaking-band-descriptors.ashx?la=en

in high-stakes assessment contexts and is grounded in empirical research and performance sampling[4]. The design of my research necessitated the application of the scale to each individual task independently rather than to the whole test, as is the operational practice. Slight modifications were therefore made to facilitate application of the rating scale to each speaking task.

One of the weaknesses of the IELTS rating scale as identified in the literature and commented on by IELTS examiners (Brown and Taylor 2006, Inoue et al 2021) is the absence of a content-oriented or a topic development criterion. Sato (2012:237) questions the reason why such non-linguistic and content-oriented criteria have yet to be defined as part of the construct definition of some general oral proficiency tests such as IELTS, arguing that 'narrowly restricting our focus to linguistic features may lead to erroneous inferences about L2 learners' ability to communicate effectively'. This was supported in Sato's (2012) empirical research showing that the content of performance of Japanese students' monologues made a 'substantive contribution' to raters' intuitive judgements of oral proficiency, leading 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 line with Sato (2012) and given the focus of the study on the effects of topic on performance, it was critical to include a content-oriented criterion that would better capture topic-related aspects of spoken performance compared to more linguistically oriented criteria. Following consideration of a number of holistic and analytic rating scales, I selected the TOEFL iBT speaking rubrics for independent tasks[5]. This scale includes a 'general description' as well as the three criteria of 'delivery', 'language use', and 'topic development'. Only the 'topic development' descriptors were used in my study.

Observation checklist

A possible approach for comparing different topics in parallel tasks is to focus on the range of language functions they elicit. A widely used instrument for evaluating speaking test tasks is the observation checklist developed by O'Sullivan, Weir and Saville (2002). The checklist includes a list of speech functions divided into *information*, *interaction*, and *interaction management* categories.

[4] 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.

[5] www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf

Originally designed for analysing language functions in paired speaking tests, the checklist has since been extensively used in various speaking assessment contexts including IELTS (Brooks 2003), Trinity Integrated Skills of English (Inoue 2013), and Test of English for Academic Purposes (Nakatsuhara 2014). A strength of the checklist is that it can be used in 'real time' to make 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 et al 2002:33).

Nakatsuhara, Inoue, Berry et al (2017), for example, compared candidates' language functions elicited in the video-conferencing delivery mode of the IST compared to its operational face-to-face mode whereas Ducasse and Brown (2011) compared functions observed in real-world classroom interactions with those elicited in the IST. In the context of my research, the observation checklist allows for a comparison of different topics (within each task type) in terms of their relative capacity to elicit a range of functions, serving to provide a source of topic validity evidence.

Score sheets

An Excel-based score sheet was designed for raters, providing the template necessary for inserting ratings for the speaking tasks, recording functions, and adding comments for each task code. Instructions for completing the score sheet were provided separately.

Gathering data

There were two main phases in the data collection. The first phase focused on collecting data from student participants and the second from raters.

Phase I data collection

In the first phase, student participants took the C-test (measure of proficiency), attended two mock IELTS interviews with myself in the role of examiner (eliciting spoken performance on 10 topics), and completed the questionnaire (measure of BK).

For the mock IELTS interviews, I randomly administered two forms of the test (either W and X or Y and Z) in succession and alternated the order to control for any order effects. It would have been preferable to administer the speaking tests on two separate occasions (Anastasi 1988), which would allow to not only evaluate the 'consistency of response to the two samples of test tasks' but to also 'measure the temporal stability of the test' (Weir and Wu 2006:169). However, administration across two occasions was likely to increase the rate of no-shows in participants. Therefore, in order to increase the chances of collecting a complete data set, I made the practical decision of administering the two speaking tests in succession.

Another concern was participant boredom or fatigue from the successive data collection procedures. I checked for this in an earlier pilot study and asked participants to comment on any boredom or fatigue but no issues were raised and in fact, participants commented that the different topics of the two tests kept them interested. Participants were also able to draw immediately on their test experiences related to all 10 topics when completing the questionnaires.

Before administering the speaking tests, I briefed participants about the format of the tests and the duration of each test and reminded them that tests would be digitally recorded. At the beginning of every recording, I identified each participant with a reference code.

Participants then completed the BK questionnaire and I was present for the duration to answer any questions in English or Farsi. Overall, the questionnaire was found to be easy to complete and straightforward though its length (80 plus questions!) was not always favourably viewed.

Phase I data processing

Speech data. The focus of my research was to compare speaking performances across different topics. This necessitated each topic – within each speaking test – to be rated separately. In Weir and Wu (2006), with a similar research design, raters were asked to listen to the full speaking test and to award separate scores for each task. Such an approach, however, runs the risk of a halo effect, where raters are more likely to assign similar scores for different topics when they are aware that they are marking the same person's performance. To address this problem, I divided each speaking test into its constituent topics using a speech editing software program and anonymised the files using a 10-digit coding system designed to make it difficult for raters to identify the same person's performance. All speech files were then divided into batches for rating.

C-tests. The C-tests were scored using two methods: a *dichotomous scoring method* where only fully correct answers were counted as correct (given a score of 1) whereas partially correct responses (either semantically or syntactically) were counted as incorrect (given a score of 0); and *a partial-credit scoring method*[6] where an incorrect response was given a score of 0, partially acceptable semantic and syntactic variants were given a score of 1, and fully correct responses were given a score of 2. I then empirically compared the scoring methods by evaluating their fit to the Rasch model (see the section 'Analysing data' for more details).

BK questionnaire. Responses to the BK questionnaire were entered into Excel for analysis.

Phase II data collection

The second phase of the research focused on scoring performances by raters. A fully crossed rating design – where all raters rate all performances – was not feasible due to the large number of speaking tasks (82 participants × 10 tasks/topics = 820). Instead, I used a 'common batch' approach 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.

I divided all speaking files into five batches: one common batch – rated by all raters – and four additional batches, rated by individual raters. Each batch contained a random allocation of performances across all tasks/topics.

Critical to the research was to ensure sufficient 'linking' in the data. The participants were connected through an overlap in tasks; all responded to tasks on Family and Leisure time. Raters were connected through an overlap in the tasks they rated (the common batch). This rating design is illustrated in Table 4.4.

[6] The argument for a partial-credit scoring system is closely linked to the Rasch measurement model score analysis. 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 is logical to assume 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.

Table 4.4 Rating matrix

Rater	Common batch	Batch (1)	Batch (2)	Batch (3)	Batch (4)
Rater 1	X	X			
Rater 2	X		X		
Rater 3	X			X	
Rater 4	X				X

Before the start of the rating processes, I scheduled one-to-one meetings with each of the raters and discussed the various instruments and procedures. I also provided them with the *IELTS Scores Explained* standard-setting DVD (UCLES 2006b) as training. This DVD explains the rating of the IST and presents benchmark performances at each band level, after which the scores for those performances are displayed. I asked raters 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. I also provided raters with a copy of the observation checklist and O'Sullivan et al's (2002) article so that they could further familiarise themselves with the instrument and its purpose.

Once rating was complete, I held a short semi-structured interview with them in order to get their insights about the rating experience. Questions related to aspects of the rating process, the rating criteria, challenges in rating, and any observations about topic-related issues and possible effects on their rating decisions.

Phase II data processing

Speaking scores. The data from raters was first screened for any missing files or incorrect data entry. The next step was to build a data set bringing together the ratings for each participant on the different topics. I used the following structure and prepared the files for analysis in FACETS (Linacre 2018b): person ID, rater ID, topic ID, scores on rating criteria.

Interview transcription. Interviews with raters were orthographically transcribed and prepared for qualitative thematic analysis.

Analysing data

The Rasch family of 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 research, and b) the construction of linear interval-level measures of speaking proficiency, BK, and general language proficiency.

Speaking score data. The speaking performance scores from each task topic were analysed using MFRM (Linacre 1989) and results were used to address the majority of RQs. Firstly, a four-facet MFRM was carried out with examinees, raters, topics, and criteria as facets (RQs i, ii, iii, iv). Secondly, a five-facet MFRM was carried out where BK of topics was conceptualised as an additional facet (RQ vi, vii). Thirdly, a bias analysis was run with MFRM (RQ viii).

C-tests. Scores from the dichotomous and partial-credit marking systems were analysed with the dichotomous and polytomous Rasch models, respectively (see the section 'Quality control' later in this chapter for more details). The results of the dichotomous marking system suggested a better fit to the Rasch model, which is why I retained these scores for further analysis. The blanks for each text in the C-test were summed and each text was analysed as a 'super-item'. This is because treating the blanks in each text as *separate items* is a violation of the assumption of local independence in Rasch analysis. The resultant linear measures were used for grouping purposes (in terms of proficiency levels) in subsequent MFRM bias analyses. The measures were also directly used as a predictor variable in a multiple-regression analysis of spoken score data.

BK questionnaire. Responses to the Likert scale questionnaire items were analysed with the partial-credit Rasch model using RUMM 2030 (Andrich, Lyne, Sheridan and Luo 2010). Linear measures of participants' BK of topics were constructed and subsequently used for grouping purposes in a five-facet MFRM and as a predictor variable in a multiple-regression analysis of spoken score data.

Multiple regression

Linear measures of BK and general language proficiency in addition to task type were used as predictor variables for topic-based spoken performance ability measures of participants (using a three-facet Rasch model) in a multiple-regression analysis with SPSS. This analysis complemented the MFRM analyses.

Descriptive statistics

The results of the functions observation checklist were analysed using descriptive statistics (frequencies) in SPSS 19.0 (IBM 2010) to address RQ v.

Thematic analysis

To address the final RQ, I conducted a qualitative thematic analysis of rater interview transcripts and any participant open comments in the questionnaires. Additionally, I used the results of the table of unexpected responses from a four-facet MFRM analysis to identify instances of speaking test performance where I hypothesised lack of BK would exert maximum influence. These performances were extracted, transcribed, and the content of speech was analysed qualitatively.

The analysis of transcriptions followed a series of qualitative steps (Ellis and Barkhuizen 2005, Strauss and Corbin 1998). In the first instance, I browsed through the transcripts for an overall impression of the data. The next step included a recursive process of note taking, sorting similar materials, labelling codes, and modifying and re-assigning of codes. This allowed for capturing of similar meanings relevant to the focus of the research. The coding process, which is an abstraction of the content of qualitative data to higher-order concepts (Ellis and Barkhuizen 2005), facilitated the reduction of large sums of text to smaller, more manageable codes. Patterns and themes then become apparent by establishing links through similarities in the coded data. I used these to provide a different qualitative perspective on the complex relations between topics and test takers' interaction with topics.

Validation of instruments

Before moving on to the main research findings (Chapters 5 and 6), I would like to report on a series of analyses carried out to validate the two main instruments that were specifically developed for this research: the C-test and the BK questionnaire. These preliminary findings provide an insight into the quality and functionality of the two instruments but are also those subsequently used in the main analyses to address some of the study's quantitative RQs (focus of Chapter 5).

C-test: pilot study and validation

Earlier in this chapter, I made a case for using C-tests as a measure of general language proficiency. In this section, I will first discuss the results of a pilot study before reporting on the findings from the main data collection.

The three versions of the C-test were piloted with a convenience sample of 203 German speakers of English as a second language. The pilot study served three main purposes: firstly, to inform the C-test scoring system to be used for the main study; secondly, to evaluate the functionality of the C-test as a measure of general language proficiency in distinguishing between persons from different proficiency levels; and thirdly, to determine a practical timeframe for administering and completing the tests. The scoring stage allowed for modifications to the answer key in specifying acceptable variants.

The C-tests were marked using two scoring systems – dichotomous and partial credit – and were then analysed with the dichotomous and polytomous Rasch models, respectively, referring to their mathematical derivation. When there are two response options for items in a scale, the dichotomous model is selected whereas for three or more options, the partial-credit model is used (Tennant and Conaghan 2007). The Rasch model is relevant to the analyses for a number of reasons:

> 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 2007:1,359).

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 is therefore 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 2009:254) while also allowing the means for evaluating the quality of the scale in the form of fit statistics (Bond and Fox 2007, Tennant and Conaghan 2007).

I used the software package RUMM 2030 (Andrich et al 2010) for the dichotomous and polytomous Rasch analyses. In order to evaluate the fit of the data to the model, I considered a number of different statistics such as the chi-squared statistics – the difference between observed values and those expected by the Rasch model – and the fit residual statistic which is based on the standardised residuals of all the responses of all the persons to an item (or criterion).

Item analysis. In the first round of 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 was examined by considering: the overall item and person fit statistics, reliability indices, and the percentage of misfitting persons and items (see the summary statistics in Table 4.5).

Interpreting results. 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. Given that this statistic approximates a standard normal deviation, it should be interpreted under the hypothesis that if the data fits 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 SD values of 0 and 1, respectively. Table 4.5 shows that the residual mean and SD 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 compared to the dichotomous model (M = −0.14). The overall person fit residual statistics are similar across the two scoring methods, both exhibiting deviations from their expected values.

Table 4.5 Summary statistics (comparison of 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	$X^2 = 1208.6$, $df = 1040$, $p = 0.00$	$X^2 = 1244.7$, $df = 1040$, $p = 0.00$
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

Note: n = 203, SD = Standard deviation, X^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-squared statistics. A large chi-squared 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 2007:1,360). The results in Table 4.5 show significant misfit of the C-test score data to the Rasch model on the basis of the chi-squared 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 a later section. At this stage, the decision was on the choice of data set to be used. A consideration of the remaining indices 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 4.5 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 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. The next step was to consider 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 2007:1,360).

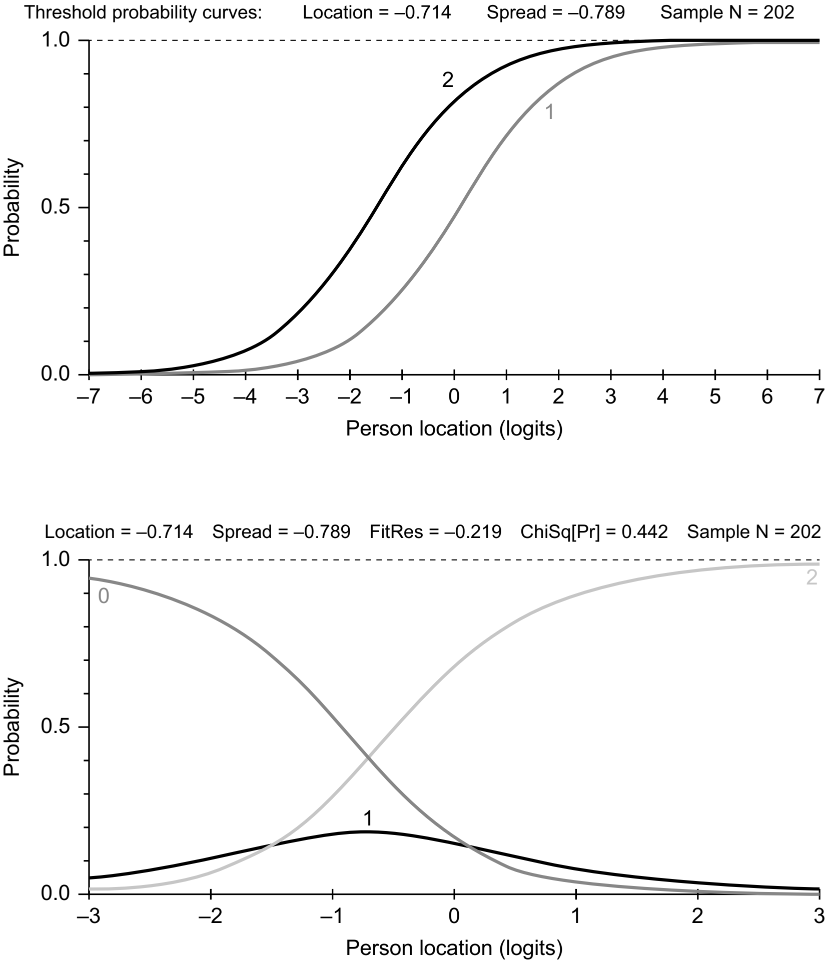
The threshold map and category probability curves (CPCs) for the 279 items from the polytomous partial-credit Rasch model (Masters 1982) were carefully studied. Results showed that the partial-credit scoring system resulted in items that displayed reversed thresholds for *all* except 28 items. These suggested that thresholds were not functioning as intended in accordance with 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 CPCs told a similar story: there was no region of the ability continuum in which a category score of 1 was most likely to be achieved. As the probability of getting a score of 0 decreased, it was the likelihood of achieving a score of 2 that increased, suggesting that the categories were not working as intended. Furthermore, in the region of person locations where one would expect a category score of 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 (an illustrative example for Item 10 is provided in Figure 4.1).

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 simpler dichotomous scoring method, which also yielded a lower percentage of misfitting persons and items.

Super-item analysis. 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). The other type of violation, *response dependence* (Marais and Andrich 2008), refers to the dependence of the response of one item to the previous item(s). The relevance of the latter response-dependence

Figure 4.1 Threshold probability curve (top) and category probability curve (bottom) for Item 10



violation in respect of C-tests is 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 279 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 PSI when items are summed, compared to when they are analysed independently (Marais and Andrich 2011).

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 parameterisation of the Rasch model for polytomous data (Masters 1982) was selected, as this parameterisation makes no assumptions regarding the equivalence of distances between thresholds in different categories, as opposed to the rating scale parameterisation (Andrich 1978), in which the distances between thresholds are held to be equal. An examination of the threshold map suggested that the distances between thresholds were not equal, which justified the selection of the partial-credit parameterisation. The map also showed no reversed thresholds.

Super-item analysis. 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 close to their expected values of 0 and 1. The chi-squared 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 dropped to 0.90 from 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-squared probabilities of individual items were also examined which, in addition to a visual inspection of the item characteristic curves (ICCs), suggested an excellent fit of the data to the Rasch model. The ICCs for all super-items were divided into the five easiest and five most difficult texts and are reproduced in Figure 4.2 and Figure 4.3. These figures illustrate that the observations (dots) are very close to their

Figure 4.2 ICCs for five easiest super-items

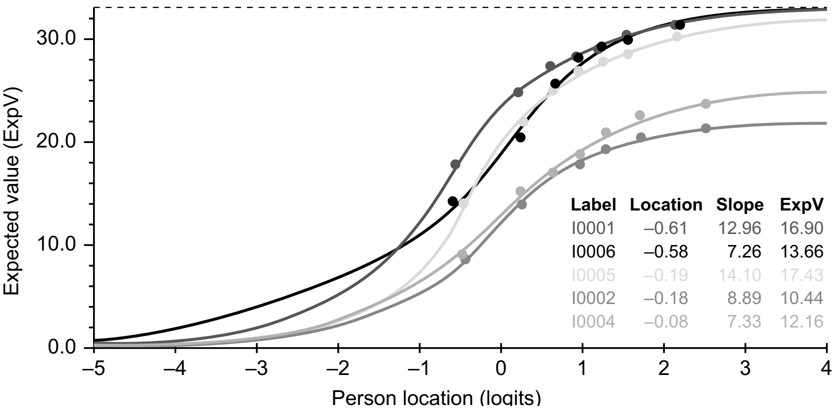
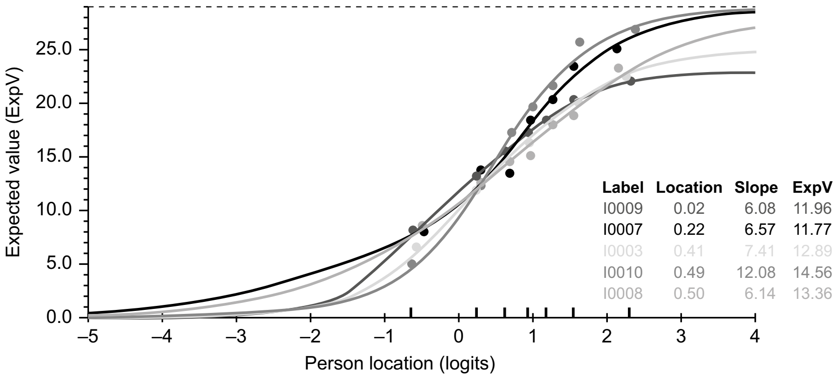


Figure 4.3 ICCs for five most difficult super-items

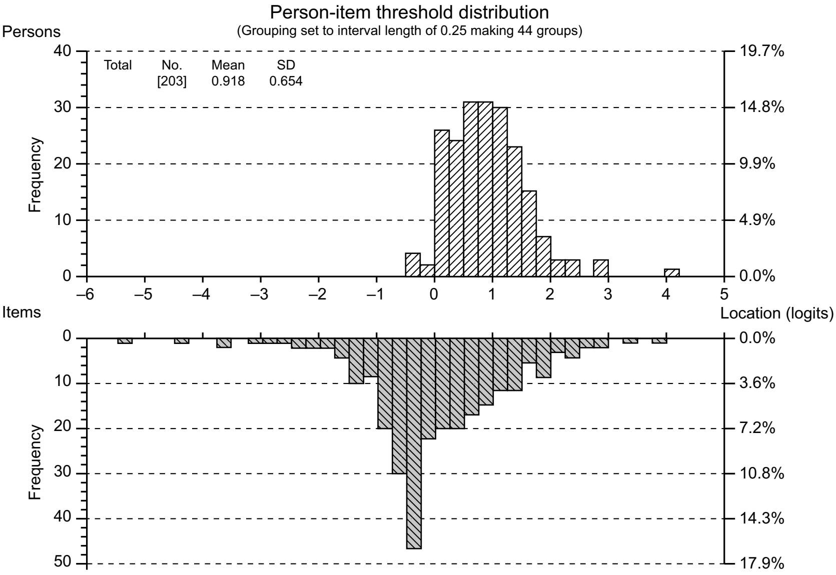


expected theoretical curves. In terms of person fit statistics, of the 203 pilot participants, only eight persons (less than 3%) had fit residuals which did not fall within the acceptable range of −2.5 to +2.5.

The last consideration was the *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 zero. Note that the mean of items in the Rasch model is arbitrarily centred at zero. 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 4.4).

The mean of the pilot participants (M = 0.91, SD = 0.65) is higher than the mean of the items. This relatively large mean value suggests that the test was

Figure 4.4 Person-item distribution map



easy for this group of test takers. 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 and provides validity evidence to support its use for the main study.

C-test: Main study results

The C-test was administered to the study's 82 participants in the main data collection phase of the research. Following the pilot study, responses were dichotomously scored and each text in the C-test was 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 4.6. 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

Table 4.6 Summary fit 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
		$X^2 = 31.1$
		$df = 40, p = 0.83$
			0.96	None

Note: n = 82, SD = Standard deviation, X^2 = Chi-squared, df = degrees of freedom, p = probability.

non-significant probability value ($p = 0.83 > 0.05$) suggest that the required property of invariance has been met. Moreover, the high PSI 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.

The super-items exhibited a range of difficulty measures from −0.84 (super-item/text 6) to 0.55 (super-item/text 10) covering 1.39 logits with all super-item fit residuals falling within the acceptable range of −2.5 to +2.5. 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-test was 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 ICCs, CPCs, and threshold probability curves (TPCs) also did not reveal any unexpected patterns or disordered thresholds.

Finally, a consideration of the targeting of the C-tests by examining the person-super-item distribution map (Figure 4.5) indicates that the test items are well targeted for the participants in the main study. The mean of the persons (M = 0.25, SD = 1.01) is located closely to the mean of the items (0) suggesting that the test is neither too easy nor too difficult for the target test takers.

To conclude, 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. The results provided validity evidence for the quality and appropriate functioning of the study's C-test as a measure of general language proficiency.

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 4.6). As explained earlier, one of the reasons why C-tests

Figure 4.5 Person-super-item distribution map

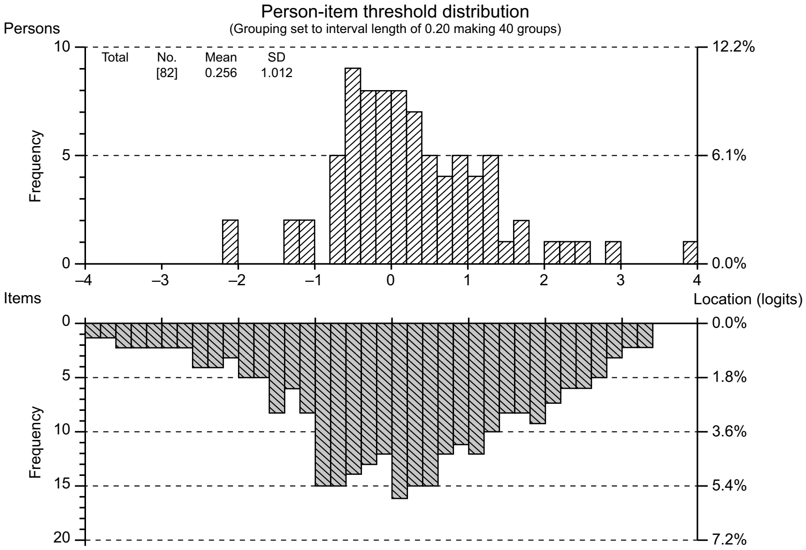
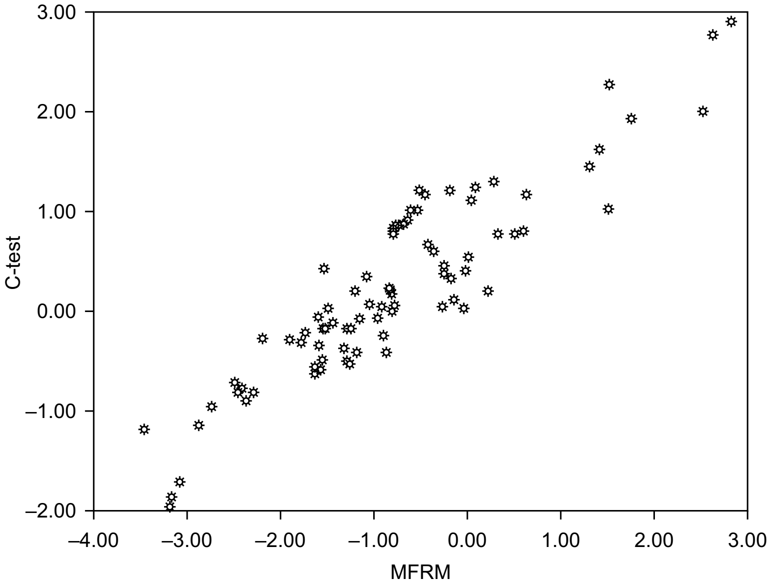


Figure 4.6 Scatterplot: speaking ability measures vs. C-test proficiency measures



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.

BK questionnaire: Validation

To elicit a measure of BK, I created a questionnaire (see the section 'Validation of instruments' earlier in the chapter) consisting of eight questions on a five-point Likert scale repeated for each topic the test takers responded to. The questionnaires were scored following the conventional assignment of integer numbers in increasing order to each response option 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.

Insights from the scoring process. The scoring process in itself provided some interesting insights into the way the respondents interacted with the questions. Most participants, for example, displayed a particular pattern in answering the questions, and that pattern was more or less repeated for familiar topics or those not particularly challenging. There was, however, a noticeable 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 was this striking shift in the pattern of responses that directed my focus to fragments of performance where lack of BK was most likely to manifest itself.

Question 5: *If I were to talk about this topic in Farsi, I would have more ideas to talk about*, in particular, elicited distinct patterns of responses. Most respondents opted for the agree or strongly agree option regardless of the topic. However, a small number of participants displayed the opposite pattern. When cross-checked 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: *I thought this was an interesting topic*. A hypothesis was that having a lot of ideas about a topic would have a positive correlation with also finding the topic interesting. For the majority of participants, there was a positive relationship between the two; however, there were also counter-examples where respondents who rated a topic as unfamiliar reported having an interest in the topic and vice versa.

Lastly, Question 8: *I performed very well on this task* required participants to assess their own performance on a given topic. Whereas some respondents provided a self-rating, there was a high proportion of respondents who were seemingly reluctant to self-assess and chose the 'undecided' option.

I would like to point out that although the questionnaire responses were primarily designed to provide a measure of topic-related BK for subsequent quantitative analyses, the above observations, which are more qualitative in nature, provided insights that were valuable in not only better understanding the quantitative results but in helping inform the types of analyses to run.

Measurement and Likert scales. Likert scales are prevalent in research in the social sciences. They are typically regarded as a soft form of data collection and subjective in nature though interestingly, their resultant scores are summed and analysed with little consideration of this subjectivity (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 is presumed to be interval (Bond and Fox 2007:102). To illustrate, consider the following two questions 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.

Both questions are eliciting attitudes towards the influence of topic on their scores; however, endorsing disagree for statement (b) might require much more of the underlying attitude, than endorsing disagree for statement (a).

Now consider another two examples related to test anxiety:

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. Yet, 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. This is 'both counterintuitive and mathematically inappropriate' (Bond and Fox 2007:101). The Rasch model instead is recommended for the analysis of questionnaire data, as it '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' (Bond and Fox 2007:103).

For these reasons, I analysed the questionnaire data with the Rasch model using RUMM 2030 (Andrich et al 2010) and evaluated the psychometric properties of the instrument and the fit of the data to the model.

Each line of data in the analysis consisted of a single participant's responses to the eight questions on each topic with 10 sets of responses available for each participant, corresponding to the topics they had taken. Each person estimate therefore reflects a person's measure of BK on a specific topic. Topics were specified as a person factor, allowing for the BK mean to be calculated for individual topics.

Evidence from the threshold map indicated unequal distances between the response categories and therefore the partial-credit parameterisation of the model (Masters 1982) was selected for analysis. There was also evidence of reversed thresholds for two of the questions – 1 and 5 – which I will address shortly.

Results. The summary statistics for the overall person and item fit residuals are presented in Table 4.7. 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-squared item-trait interaction value ($\chi^2 = 428.88$, $df = 72$) and a significant probability value of $p = 0.00 < 0.01$.

Table 4.7 Summary statistics for BK questionnaire

Overall fit residuals (items)	Overall fit residuals (persons)	Total item-trait interaction	Person separation index	% of misfitting items/persons (fit values > |2.5|)

Mean = −0.78
SD = 6.55
	Mean = −0.42
	SD = 1.28
		$X^2 = 428.88$
		$df = 72, p = 0.00$
			0.90	1.7%

Note: n = 809, SD = Standard deviation, X^2 = Chi-squared, df = degrees of freedom, p = probability.

The fit residual statistics for persons are M = –0.42 and SD = 1.28, which despite some deviations from their expected values, are close enough to not raise serious concerns regarding misfit. The PSI (0.90) and its traditional counterpart, Cronbach Alpha ($\alpha = 0.92$), suggest the questionnaire – although short in length – can reliably distinguish between persons with lower and higher BK levels.

The individual item difficulty estimates showed a range of difficulty measures for the different questionnaire stems covering a range of 1.05 logits. In line with the observations from the questionnaire scoring process, Item 8 (self-assessment of performance) was the most difficult item for respondents to endorse (+0.51 logits) and Item 1, which required respondents to rate their familiarity with a topic, the easiest (−0.54 logits).

Different fit indices were taken into account in flagging items that displayed misfit. Table 4.8 shows that the residual values for Items 1 to 4 (shaded in dark grey) are well below the minimum criterion value of −2.5, indicating overfit to the model. On the other hand, Items 5 and 7 (shaded in light grey) 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 BK. 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 (Eckes 2009, Myford and Wolfe 2003).

Misfit to the model is also evidenced in chi-squared statistics with probability values below $p = 0.05$ (identified with *). 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 6 and 8, displayed significant misfit.

Table 4.8 Questionnaire item fit statistics

Item	Measure	SE	Residual	X^2	*df*	*P*
1	−0.54	0.05	−5.33	28.73	9	0.00*
2	−0.40	0.05	−6.99	53.08	9	0.00*
3	−0.11	0.05	−6.92	50.73	9	0.00*
4	−0.13	0.05	−5.32	35.92	9	0.00*
5	0.44	0.05	11.21	186.56	9	0.00*
6	0.43	0.05	1.85	6.42	9	0.69
7	−0.19	0.05	4.72	39.55	9	0.00*
8	0.51	0.05	0.49	27.87	9	0.83

Note: SE = Standard error, X^2 = Chi-squared, df = degrees of freedom, p = probability.

Given the misfit, it was important to further evaluate the data from a diagnostic perspective to decide on possible remedial actions. One possible source of misfit is disordered thresholds, or categories not functioning as intended. As mentioned earlier, two of the items, 1 and 5, displayed reversed thresholds. The threshold and CPCs for these two items were therefore scrutinised more closely (see Figure 4.7 and Figure 4.8).

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 TPC

Figure 4.7 TPCs (left) and CPCs (right) for Item 1

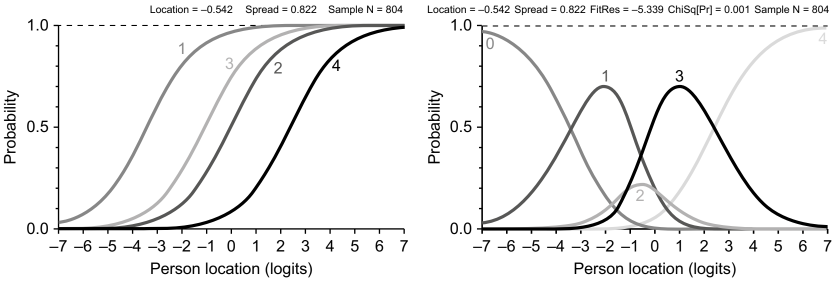
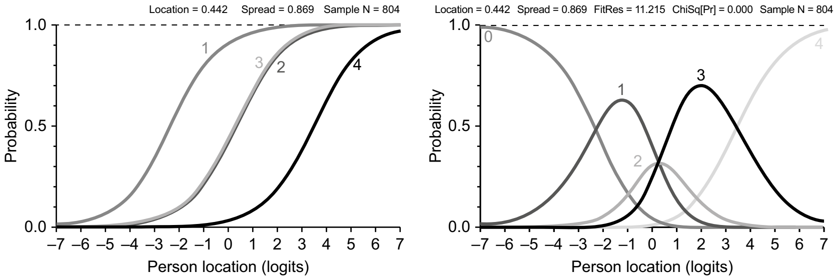


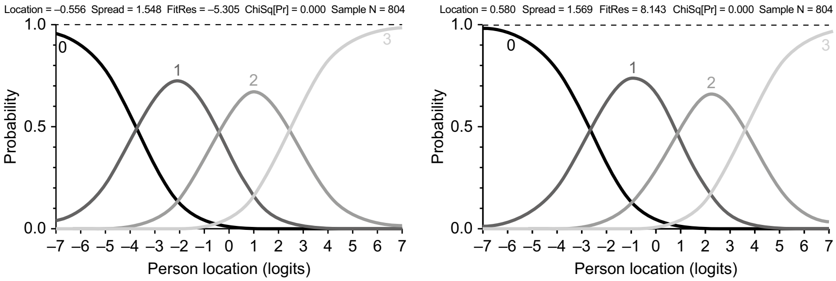
Figure 4.8 TPCs (left) and CPCs (right) for Item 5



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 with 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, suggesting that there is little or no discrimination between their respective categories.

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. Concerns with designating the undecided or 'not sure' category in the middle of a response scale have been raised by researchers. Andrich, De Jong and Sheridan (1997:66–67), for example, on the basis of evidence from their study of teacher attitude questionnaire data observe that 'the scoring of the not sure category as if it is operating in the middle of the other categories is not tenable'. For the current study, it was clear that the middle category was not functioning as intended for two of the questionnaire items. I therefore followed the approach suggested by Andrich et al (1997:70) for dealing with this issue and collapsed the disordered categories. The CPCs for Items 1 and 5 following this remedial action can be viewed in Figure 4.9. The categories are now ordered and the threshold map no longer exhibited disordered thresholds. Moreover, the fit residual statistics for both items showed improvements following this remedial action.

Figure 4.9 CPCs for Items 1 (left) and 5 (right) after remedial action



Another source of misfit that has been identified in the literature is that of differential item functioning (DIF), 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 2007:6).

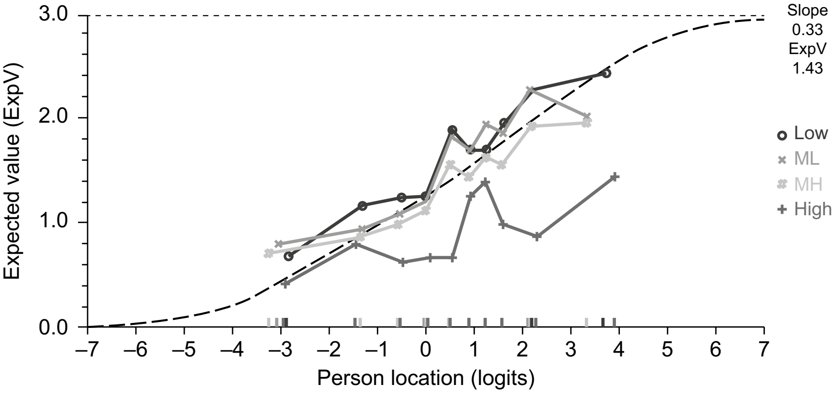
Given my initial observations from the questionnaire scoring process (in relation to Item 5), I hypothesised that the item might be displaying DIF for persons of different proficiency levels. To empirically test this out, I specified a person factor of *proficiency group* and used the C-test ability estimates to divide participants into four proficiency groups (low, medium-low, medium-high, and high). Next, I carried out a DIF analysis; results showed significant

DIF for Items 5 and 8 although only Item 5 was flagged once the Bonferroni adjustment was used.

The graphical display of Item 5's ICC, in which different proficiency levels are plotted over (Figure 4.10), show an clear contrast between the way the high-proficiency group responds to this question compared to the other groups. We can better understand this by re-visiting the questionnaire stem for Question 5: *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 L2 to their language proficiency (or lack thereof), whereas for the high-proficiency groups, language is no longer a barrier. Therefore, if the individuals in the latter group have enough ideas, they are likely to be able to express them in their L2, and if they are unable to do so, it is an indication that BK-related problems persist in their first language. This is illustrated in the ICC where the observations for the high-proficiency group are well below the expectations of the model.

Figure 4.10 ICC for Item 5; grouped by proficiency level (ML = medium-low, MH = medium-high)



The above analysis suggested that Item 5 was not working as intended for the high-proficiency group. I therefore applied an item-split or *item resolve* method[7] (Andrich and Hagquist 2012) as a remedial action; however, results showed improvements only for the low-proficiency group and inconsistencies for the remaining groups, with the high-proficiency group displaying significant misfit to the model with almost no discrimination along

[7] Resolving an item refers to creating an item specific to different levels of the person factor; for example, one item for the low-proficiency group, one item for the medium-low-proficiency group, and so on.

the continuum. Given the ineffectiveness of this approach on the one hand, and the large contribution of this item to overall misfit of data to the Rasch model on the other, I decided to remove the item from subsequent analyses and to re-examine the ICCs. Results showed that for the remaining items, observations were close to the theoretical curve and model expectations. Note also that in examining misfit we are looking at deviations from an ideal and that it is the 'practical utility' of the model which should be taken into account:

> 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 (Eckes 2009:27).

Taken together, 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 BK questionnaire provided strong evidence for its validity and support its use as an instrument which can reliably distinguish between higher and lower levels of topic-related BK. These BK estimates were therefore used to address the study's RQs.

Topic range

The estimated BK measures of persons on each topic (in logits) also allowed for an examination of the relative familiarity of topics by considering the BK means of persons for each topic. A higher mean indicates topics for which participants, as a group, reported higher levels of BK whereas a lower topic mean indicates low levels of familiarity with a topic.

Table 4.9 shows the topics, the number of persons who rated their BK of each individual topic, their mean (in logits), and SD values. The topics are divided by task type and serially ordered. Results show differences in BK means on different topics; for example, topic B.3 (Describe someone in your family) is associated with the highest BK mean whereas the average BK estimates on Topic C.6 (Genetic research) is the lowest across all topics. The results of an analysis of variance (ANOVA) indicate that observed differences are statistically significant [F (17,787) = 14.88, $p < 0.001$] thus confirming that the study has been successful in selecting a range of topics for which participants have varying levels of BK.

Table 4.9 BK means for each topic

Topic reference	Topic name	N	Mean	SD
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 in contact	40	1.83	1.79
A.6	Dancing	40	1.46	2.32
B.1	Describe a friend	42	2.09	1.61
B.2	Describe a river, lake or sea	42	1.01	2.52
B.3	Describe someone in your family	40	2.75	1.93
B.4	Describe an important 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 leisure activities	42	0.24	2.09
C.4	The economic importance of rivers, lakes and the sea	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	Choices in everyday life	40	1.01	2.32

Quality control

The quality and rigour of research findings are directly influenced by the quality of the instruments used for data collection. In this chapter I have discussed the methodology used in my research and provided details of the study's instruments and various data collection and analysis procedures (for schematic representations of the different stages of data collection and analyses as well as the various instruments used at each stage see Figure 4.11 and Figure 4.12, respectively). I also evaluated the psychometric properties of the C-test and BK questionnaire using the Rasch family of models by bringing together different pieces of evidence such as fit statistics, reliability indices, and ICCs as well as taking remedial action where necessary.

Taken together, the analyses lend strong support to the appropriateness of these instruments for their intended purposes and for the transformation of raw scores to Rasch-based interval-level measures of general language proficiency and topic-related BK levels. Having established their quality, I will be drawing on these measures to examine the effects of topic and BK of topic on speaking performance in the next chapter.

Figure 4.11 Data collection

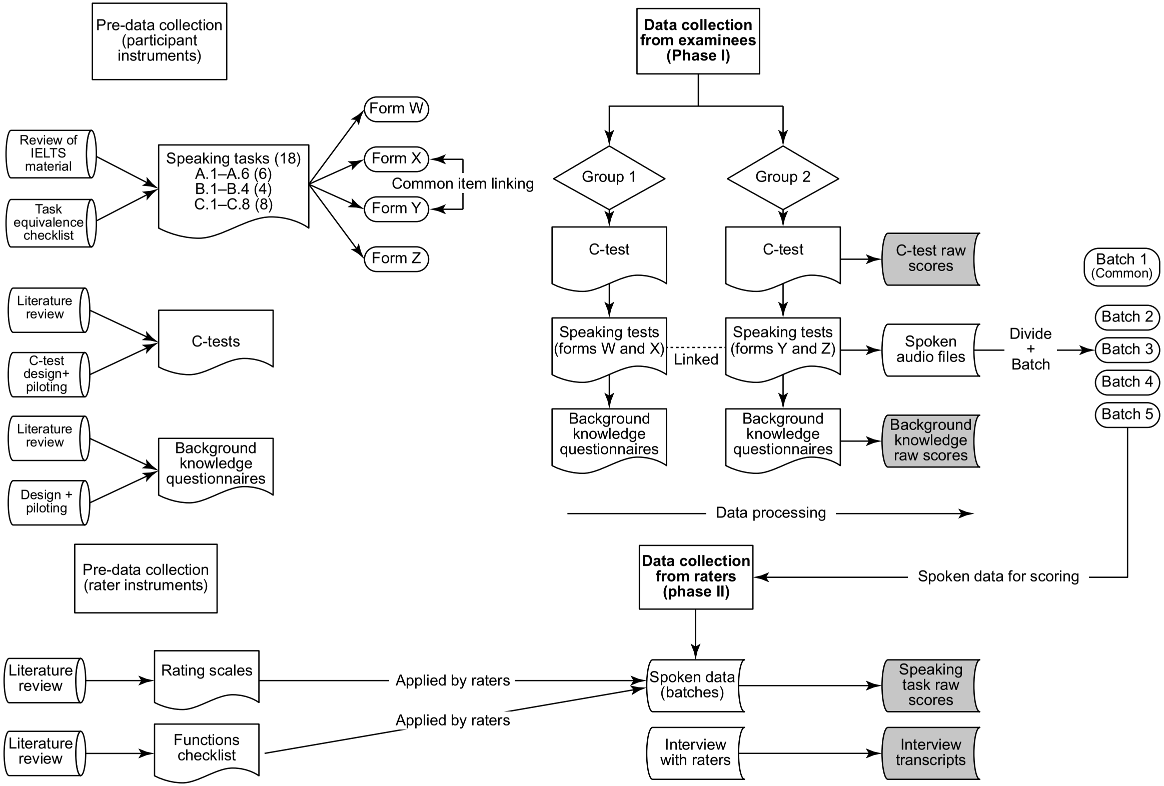
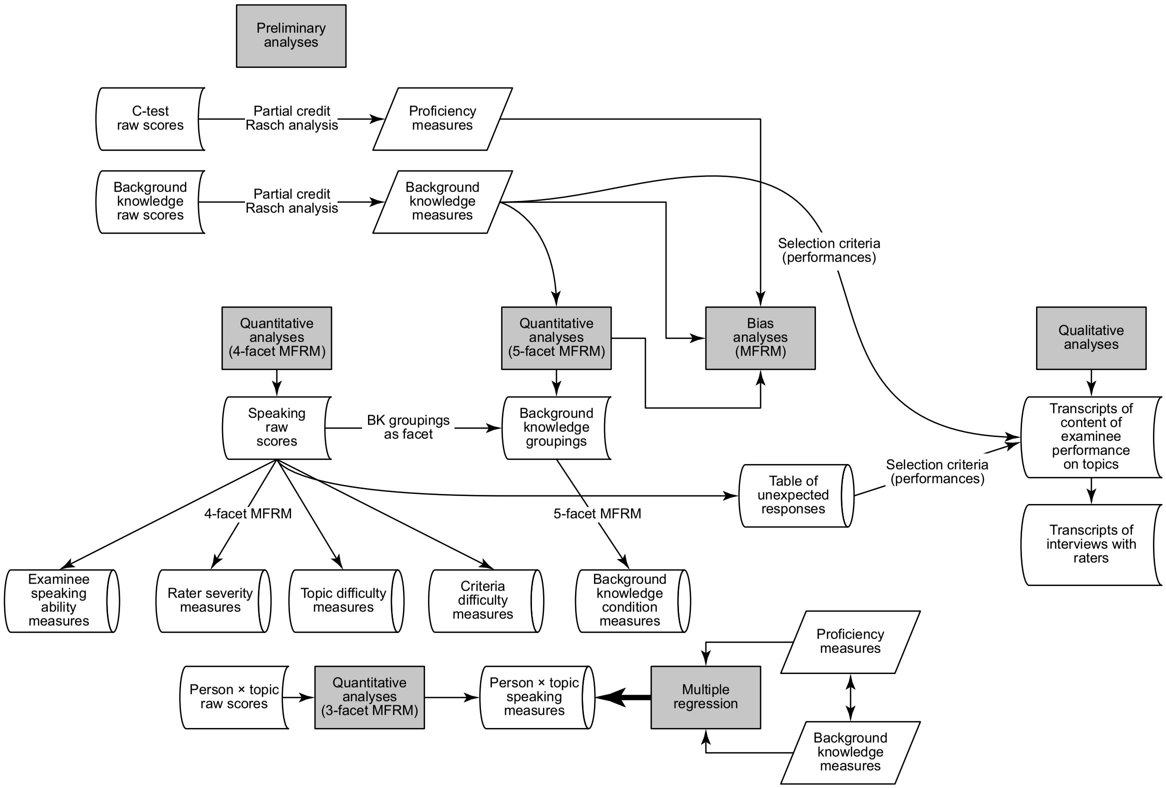


Figure 4.12 Data analysis


Preliminary analyses
C-test raw scores
Partial credit Rasch analysis
Proficiency measures
Background knowledge raw scores
Partial credit Rasch analysis
Background knowledge measures
Selection criteria (performances)
Quantitative analyses (4-facet MFRM)
Quantitative analyses (5-facet MFRM)
Bias analyses (MFRM)
Qualitative analyses
Speaking raw scores
BK groupings as facet
Background knowledge groupings
Transcripts of content of examinee performance on topics
Table of unexpected responses
Selection criteria (performances)
4-facet MFRM
5-facet MFRM
Transcripts of interviews with raters
Examinee speaking ability measures
Rater severity measures
Topic difficulty measures
Criteria difficulty measures
Background knowledge condition measures
Proficiency measures
Person × topic raw scores
Quantitative analyses (3-facet MFRM)
Person × topic speaking measures
Multiple regression
Background knowledge measures

5 Does choice of topic matter? *A quantitative perspective*

The focus of this volume has been on an examination of the extent to which L2 spoken performance is affected by the variables of interest, namely *topic* and test takers' *background knowledge of topic* (BK). To address this, we will look at the study's research findings and consider the influence of topic from different angles: a measurement angle (focus of this chapter) and a qualitative angle (focus of the next chapter).

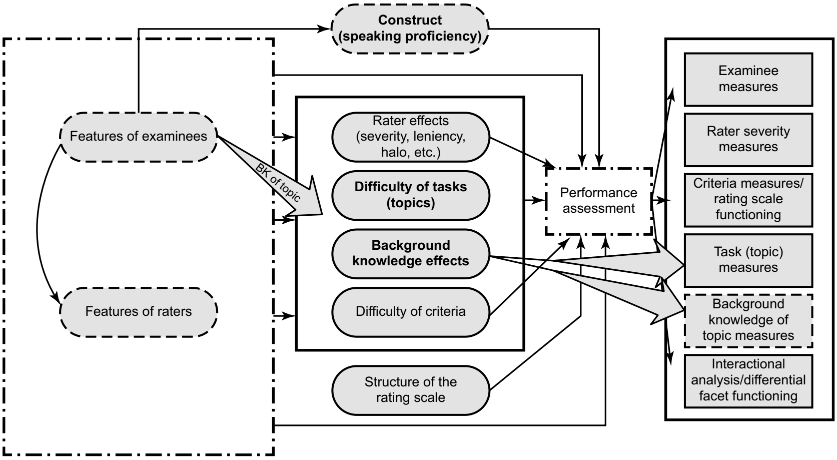
As discussed in Chapter 3, the assessment of speaking is a complex process, which is influenced by a number of factors as well as the interactions between them (Eckes 2009, McNamara 1996) and therefore, to better understand the facets of interest, they need to be embedded within this larger picture. To this end, I will start the chapter with an introduction to a framework that helps contextualise the study.

A conceptual-psychometric framework

A framework can facilitate the systematic reporting of the MFRM results of the study. Figure 5.1 is adapted from the 'conceptual psychometric framework' (Eckes 2009:11) and provides a schematic view of the most relevant factors that can influence spoken performance in my research. The original framework was designed for the assessment of writing and 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 than can be summarized in a diagram, and the factors coming into play are diverse at any given moment'.

The central box in the diagram represents 'proximal' factors – those with an immediate impact on scores – the most important of which is the construct of interest, that is, the spoken ability of examinees. Other factors such as rater effects, variability in task difficulty, and rating criteria contribute to a systematic source of measurement error (Eckes 2009). These are distinguished from 'distal' factors (in the box on the left) which exert 'additional influence on the ratings, albeit usually in a more indirect and diffuse way' (Eckes 2009:10). Examples are individual characteristics of examinees and raters. The original framework (Eckes 2009) includes the characteristics of the testing situations such as the technical and physical environment, which are particularly important in large-scale commercial testing of speaking in

Figure 5.1 A conceptual-psychometric framework (adapted from Eckes 2009:11)


Construct (speaking proficiency)
Features of examinees
BK of topic
Features of raters
Rater effects (severity, leniency, halo, etc.)
Difficulty of tasks (topics)
Background knowledge effects
Difficulty of criteria
Structure of the rating scale
Performance assessment
Examinee measures
Rater severity measures
Criteria measures/ rating scale functioning
Task (topic) measures
Background knowledge of topic measures
Interactional analysis/differential facet functioning

relation to aspects of test fairness and equity for candidature. In my research study, however, these features were of secondary importance and therefore excluded. Other factors such as the characteristics of the interlocutor or examiner (O'Sullivan 2000) were also excluded, as these were controlled for to a large extent by having the same interlocutor administer all speaking tests and following a strict examiner script. On the other hand, BK of topic, which falls under 'distal' factors as a test taker characteristic in the Eckes (2009) model, is conceptualised as a *proximal* factor in my study, hypothesised to exert a direct impact on spoken performance. The framework visualises the interactions within and between the two categories of factors as connecting arrows. Lastly, the box on the right of the diagram identifies the main types of measurement reports that are produced from an MFRM analysis: 'MFRM modelling 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).

In line with the conceptual-psychometric framework (Eckes 2009), we will look at the MFRM results of the study with (a) four proximal facets, (b) five proximal facets, and (c) the interaction between facets. The results will be drawn on to address the topic validity of the speaking assessment under examination from the perspective of scores.

The overall picture: Facets of assessment

Before delving into the influence of topics on performance, let us first consider the different facets of this specific performance context. This serves not only as a quality control check for the various aspects of the speaking test but also to better understand and contextualise topics within the overall picture of the assessment setting.

In the first analysis, the following four proximal facets (and elements within each facet) were identified with speaking task topics explicitly parameterised as a facet:

- examinee facet (81[1] participant elements)
- rater facet (four rater elements)
- rating scale (five criteria elements)
- topics (18 topic elements).

MFRM generates a range of statistics, a careful consideration of which allows for an understanding of the different facets, facet elements, and the interactions between them from the perspective of scores.

Each statistic provides a different piece of information regarding the fit of the data to the Rasch model 'similar to viewing something from a different angle' (Tennant and Conaghan 2007:1,360). These include parameter estimates for each facet and corresponding reliability indices, the separation statistics which are useful for summarising observations and drawing inferences about group trends, and the separation indices and strata which estimate the number of statistically distinguishable levels and their associated reliability (Linacre 2018a). In addition to group-level statistics, FACETS (Linacre 2018b) also generates a series of 'fit statistics', which 'enable the diagnosis of aberrant observations and idiosyncratic elements' (Linacre 2018a:14) within each facet. I will discuss the analyses and the findings for each facet before turning to the study's RQs.

In the first analysis, the four facets were mapped onto a common interval-level scale known as the logit scale (log-odd units) and visually represented in the vertical map in Figure 5.2. The logit scale 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 illustrates, in a graphical form, the calibrations for all examinees, raters, topics, rating criteria, and scale categories, and the logit scale serves as a single frame of reference for interpreting the results of the analyses.

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. Should the data fit the model, then this logit scale would constitute an interval-level scale necessary for measurement (Tennant and Conaghan 2007:1,358).

[1] The audio file for one participant was corrupt and had to be removed from remaining analyses.

Figure 5.2 Four-facet MFRM map

```
+----------------------------------------------------------------------+
| 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  +     +     |
|     | *       |       |        |       |     |     |     |     |     |
|     |         |       |        |       |     |     |     | --- | --- |
|     |         |       |        |       |     |     |     |     |     |
|     |         |       |        |       |     |     |     |     |     |
|     |         |       |        |       |     |     |     |  7  |     |
|   2 + *       +       +        +       +     +     +     +     +     |
|     | **      |       |        |       |     |     |     |     |     |
|     | *       |       |        |       |     |     | --- |     |  4  |
|     | *       |       |        |       | --- | --- |     | --- |     |
|     |         |       |        |       |     |     |     |     |     |
|     |         |       |        |       |     |     |     |     |     |
|   1 +         +       +        +       +     +     +     +     +     |
|     | **      |       |        |       |     |     |     |     | --- |
|     | *       |       |        |       |     |     |  7  |     |     |
|     | **      |       |        | P     |  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 + *       +       +        +       + --- + --- + --- +     +     |
|     | *       |       |        |       |     |     |     |     |     |
|     | ****    |       |        |       |     |     |     | --- |     |
|     |         |       |        |       |     |     |     |     |     |
|     | **      |       |        |       |     |     |     |     | --- |
|     |         |       |        |       |  5  |  5  |  5  |     |     |
|  -3 + **      +       +        +       +     +     +     +     +     |
|     | *       |       |        |       |     |     |     |     |     |
|     | *       |       |        |       |     |     |     |     |     |
|     |         |       |        |       |     |     |     |     |     |
|     |         |       |        |       |     |     |     |     |     |
|     |         |       |        |       | --- | --- |     |  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      | FC  | LR  | GA  | P   | TD  |
+----------------------------------------------------------------------+
|Mean | -0.63   | 0.00  |  0.00  | 0.00                                |
|SD   |  1.63   | 0.45  |  0.22  | 0.66                                |
+----------------------------------------------------------------------+
```

Note: Each star (*) in the second column represents one examinee.
Measr = Measure, Cri = Criteria, FC = Fluency and Coherence, LR = Lexical Resource, GA = Grammatical Range and Accuracy, P = Pronunciation, TD = Topic Development.

The **second 'examinee' column** displays the examinee speaking proficiency estimates where each star (*) denotes a test taker. The (+) sign next to the examinee label of the column indicates 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. This allows us to examine the distribution of the examinees. Given that the difficulty of the rating criteria is centred at zero, we can see that most examinees are clustered around 0 to −2 logits on the basis of their scores on the rating scales across criteria. 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 column label 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 they assigned to examinees. We can see that Rater 3 (R3) is strikingly more lenient than the other raters, and R2 is the most severe of the four.

The **fourth 'topic' column** displays the 18 topics used in the research (see Appendix B for the full tasks/topics). Letters A, B and C denote the type of task: Interview or Information Exchange (A), Individual Long Turn (B), and Two-way Discussion (C) whereas the numbers specify the topics within each task type. Similar to the 'rater' facet, the 'topic' facet is negatively oriented with higher measures corresponding to higher task difficulty. We can see that speaking tasks C.4 and C.6 are amongst the most difficult topics and Topics B.4 and A.6 are the easiest. Topics C.7 and A.4 are estimated to be of average difficulty.

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 IELTS criteria of Fluency and Coherence (FC), Lexical Resource (LR), Grammatical Range and Accuracy (GA), and Pronunciation (P) exhibit similar difficulty levels. On the other hand, Topic Development (TD) – a criterion specifically added for the purpose of this research – is considerably easier, judging by its notable distance from the other criteria. In other words, TD was the easiest criterion for examinees to get a high score on.

Lastly, the remaining five columns to the right of the map display the nine-category scale for the IELTS speaking criteria (FC, LR, GA and P) and the five-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 categories. The horizontal dashed lines in these columns indicate the category thresholds, that is, the point at which an examinee with an average expected score has a 50% probability of being assigned to one of the two adjacent categories. 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 SD of the distribution of measures (in logits) for the four facets are displayed. 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 2018a).

In the next sections, we will go through the MFRM results for each facet in order to first establish the parameters of the speaking assessment context before discussing the topic facet in more detail.

The examinee facet

The examinee measurement report (n = 81) suggested a range of ability levels from −3.37 logits (Examinee 65) to +3.44 (Examinee 40) spanning 6.81 logits. This range is illustrated visually in the wide distribution of examinees in the facet map (Figure 5.2). For each examinee, FACETS reports an 'observed' average and a 'Fair-M' average. The former is a given examinee's scores summed across tasks and raters, divided by the observed count (Linacre 2018a). The Fair-M average, on the other hand, is the observed average, but adjusted for differences in other facets. This is a useful statistic which is, in effect, the converted form of the Rasch measure into the raw-score metric of the original scales (Eckes 2009, Linacre 2018a). Differences in raw scores may reflect differences in speaking proficiency but they may also relate to differences in the severity of raters or difficulty of tasks assigned to different examinees. The Fair-M average addresses this issue by disentangling the influence of other facet elements from examinee speaking proficiency measures (Eckes 2009) and thus allowing for 'fair' comparisons to be made. Table 5.1 presents the observed average, Fair-M average and SDs for the whole sample for (a) all five criteria, (b) the nine-band IELTS scale, and (c) the five-level TD scale. Results show that the values for observed averages across analyses are close to the adjusted Fair-M averages. The influence of other facets on the examinee estimates is therefore minimal.

The other advantage of the Fair-M average is the reporting of the logit measure in terms of the original scale; this greatly facilitates data interpretation. To illustrate, the Fair-M average and SD for the sample of examinees in this study is Fair-M = 5.95, SD = 0.79 when the data is analysed for the IELTS criteria. The reporting of the Fair-M average on the same metric as the IELTS nine-band scale allows for a direct comparison of this statistic with other research and/or publicly available IELTS data. The mean band score reported for Iran in the 2011 IELTS Test Taker Performance Report was M = 6.3, which suggests that the participants' speaking scores in my study were, on average, only less than half a band level lower ($0.35 < 0.5$) than the Iranian test-taking population in that year. This can be taken as evidence that, despite the small sample size, the study has been successful in selecting participants that not only exhibit a range of speaking ability levels but that are representative of the IELTS test-taking population in terms of their average levels.

Table 5.1 Examinee statistics (observed and Fair-M averages)

Analysis	**Observed average**		**Fair-M average**	
	Mean	**SD**	**Mean**	**SD**
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 scale range is 1–5.

Now let's turn our attention to a series of group-level statistics. The separation index (denoted by G) is the ratio of the true SD to the average measurement error and estimates the number of statistically distinguishable performance levels (Linacre 2018a). 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 of examinees ($G = 8.64$) suggests that there are at least eight 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 modelled as extreme "true" levels of performance' (Linacre 2018a: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 a performance assessment context such as the one in this study, extreme scores are likely to present persons of relatively high or low speaking abilities. Given the small sample size, it is also not surprising that there are only a few persons at the two extremes. The strata statistic therefore provides a more accurate measure of the spread of examinee ability. The 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 define reliably the facet' (Eckes 2009:20). This value gives 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). As discussed earlier, the interpretation of this index is facet-dependent: for examinees, a high separation reliability value (close to 1) is preferable. After all, a good test instrument should be able to separate people into different levels of ability. In contrast, for a facet where elements should be similar in measures (e.g. for raters), a low value (close to 0) is desirable. The high separation reliability value for the examinees in the study (0.99) suggests 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 measurement report also includes a 'homogeneity statistic' (Eckes 2009:19): an overall chi-squared 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 2018a). For the examinee facet, it tests the null hypothesis that examinee ability measures are the same, once measurement error has been allowed for. Results ($X^2 = 6043.5$, $df = 80$, $p = 0.00 < 0.05$) suggest that the null hypothesis has been rejected and that variability observed in speaking ability measures is statistically significant.

Moving away from group-level statistics, individual fit statistics were considered to evaluate the extent of fit (or misfit) of the data to the Rasch model. FACETS reports both infit and outfit mean square statistics. Infit is broadly viewed as more important than outfit in evaluating the fitness of the data to the model, as it is less sensitive to outliers or unexpected ratings but rather, more affected by the accumulation of unexpected ratings (Eckes 2009, Myford and Wolfe 2004). Values below 1 are considered to be 'overfitting' the model and too predictable, whereas values above 1 are considered to be 'underfitting' and too unpredictable (Linacre 2018a), with the latter generally raising more cause for concern (Eckes 2009, Linacre 2018a).

There are no hard-and-fast rules in setting quality control limits for the infit index (Myford and Wolfe 2000). Eckes (2009) advises for a consideration of the stakes of the test and the purposes of assessment in deciding on an acceptable range. Linacre (2018a: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 (Bond and Fox 2007, McNamara 1996).

For this research and in line with Linacre (2018a), I adopted lower and upper control limits of 0.5 and 1.5, respectively, for the infit mean square index. Table 5.2 shows the examinees displaying misfit; only five out of 81 examinees (6%) exhibited infit indices outside of the acceptable range with Examinee 49 displaying overfit (infit < 0.5) and the remaining four showing underfit (infit > 1.5). The values for Examinees 49, 5, and 15 are quite close to the acceptable range and therefore do not raise serious concerns. On the other hand, Examinees 1 and 8 – amongst the top five highest-ability examinees in the sample – display fit statistics much higher than model expectations.

Table 5.2 Examinees displaying misfit

E ID	Measure	Model SE	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

Note: E = Examinee, SE = Standard error, MnSq = Mean square.

To find an explanation for these observations, I examined the table of unexpected responses. As its title suggests, this table produces a list of responses with associated standard residual values above ±3 for any combination of facet elements. Interestingly, both Examinees 1 and 8 – high-ability participants – were flagged in this table for exhibiting scores much lower than expected on certain topics for the TD criterion, the easiest of all criteria. I will come back to this finding later in the chapter.

Overall, the small number of misfitting examinees (6% with only 2.4% exhibiting large underfit), and the results of different outputs for the examinee facet suggest that despite its small sample size, the study has been successful in selecting participants who exhibit a range of statistically distinct speaking ability levels and are representative of the IELTS test-taking population in terms of their average levels.

The rater facet

The review of the literature outlined some of the ways in which raters can introduce measurement error to performance assessment, the most problematic of which is variability in rater severity (Eckes 2009, McNamara 1996). Other rater-related systematic tendencies included extremism, central tendency, halo, and bias effects (Linacre 2018a, Myford and Wolfe 2003, 2004). In the next sections I will examine some of these rater tendencies in more detail.

Note that in reporting findings, I will be consistently reporting two sets of results from: (a) analyses with all five criteria (henceforth FullA), and (b) analyses with the IELTS criteria only where the TD criterion has been removed (henceforth IELTSA). To remind the reader, the TD criterion was included in this study as a means of isolating the effects of topic on performance. Running the separate analyses serves two purposes: firstly, it facilitates the linking of the findings to the IELTS speaking context and other research studies; and secondly, it allows for an exploration of the TD criterion and its influence in relation to the other criteria.

FACETS (Linacre 2018b) directly parameterises rater severity as a facet and reports severity estimates for each rater. The measurement reports for the raters in this study are presented in Table 5.3 (FullA) and Table 5.4 (IELTSA) in increasing order of severity. In both analyses, Rater 3 (R3) is the most lenient rater ($R3_{FullA} = -0.65$, $R3_{IELTSA} = -1.06$) and R2 is the harshest ($R2_{FullA} = +0.34$, $R2_{IELTSA} = +0.45$), with a logit difference of 0.99 and 1.51, respectively. The standard error (SE) for all rater estimates is SE = 0.3 in FullA and SE = 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 5.3 Rater measurement report (FullA)

Rater (R) ID	Observed average	Fair-M average	Measure	Model SE	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
SD	0.2	0.26	0.45	0.00	0.16	0.14

Model, Sample: RMSE .03, Adj (True) SD: .45, Separation: 12.14, Strata: 16.52, Reliability (not inter-rater): .99
Model, Fixed (all same) chi-square: 586.9, *df*: 3, Significance (probability): .00
Inter-rater agreement opportunities: 5915, Exact agreements: 2173 = 36.7%, Expected: 2187.2 = 37.0%

Note: SE = Standard error, MnSq = Mean square, SD = Standard deviation, RMSE = Root Mean Square Standard Error, Adj (True) SD = 'True' sample standard deviation of the estimates after adjusting for measurement error, df = degrees of freedom.

Table 5.4 Rater measurement report (IELTSA)

Rater (R) ID	Observed average	Fair-M average	Measure	Model SE	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
SD	0.3	0.34	0.71	0.00	0.19	0.20

Model, Sample: RMSE .04, Adj (True) SD: .71, Separation: 18.32, Strata: 24.76, Reliability (not inter-rater): 1.00
Model, Fixed (all same) chi-square: 972.1 *df*: 3, Significance (probability): .00
Inter-rater agreement opportunities: 4732, Exact agreements: 1683 = 35.6%, Expected: 1791.3 = 37.9%

The facet map in Figure 5.2 already gave us a general sense of the extent to which raters differ in their severity. 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 logits) 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 IELTSA is 8.49, which is 5.6 times the spread of rater severity (1.51). These distributions generally suggest that differences in rater severity are unlikely to exert a large influence on examinee scores (Myford and Wolfe 2000).

To further investigate the potential impact of rater severity, a useful statistic to consider is the Fair-M average for different raters. As explained earlier, the Fair-M average is a powerful statistic which disentangles the severity of the rater from the proficiency of the examinees they 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 band scores apart. A larger difference of 0.74 band scores is observed for IELTSA. We can contextualise this value; in IELTS, the smallest unit that can have a practical effect on examinee scores is half a band (0.5 band scores). The difference in rater severity exceeds this (0.74 > 0.50), which means that examinees in this sample might be given scores that are half a band apart depending on the rater. Such differences in severity are documented in other research. In fact, the trained IELTS examiners in O’Sullivan and Rignall (2007:456) exhibited much larger differences in severity of 1.2 band scores compared to 0.74 in this research. Other studies with the IST have reported severity differences closer to half a band; for example, 0.36 of a band in Nakatsuhara, Inoue and Taylor (2017), 0.52 of a band amongst the 10 examiners in Nakatsuhara, Inoue, Berry et al (2017) and 0.76 of a band amongst the eight examiners in Berry et al (2018).

A more pressing matter is whether the raters in the study are 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, summarised over examinees and criteria' (Eckes 2009:16). The rater fit statistics in Table 5.3 and Table 5.4 are all within the acceptable range of 0.5 to 1.5 as well as more stringent lower and upper control limits of 0.7 to 1.3. These fit indices demonstrate that regardless of differences in harshness and leniency, the raters in this study are consistent within themselves and fit the model. These findings align closely with the literature on systematic rater effects: raters tend to display high levels of self-consistency in rating but nevertheless exert a significant influence on examinee scores owing to differences in severity levels (McNamara 1996). Taken together, the rater results suggest that with minimal training (with the *IELTS Scores Explained* DVD) and without any standardisation procedures, the raters in my study were not only consistent, but also exhibited a narrower range of severity levels than trained IELTS examiners in some other studies in the literature.

We can also frame this discussion within Linacre's (2018a: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 which category the raters belong to, 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. An exact agreement percentage of close to 90% is indicative of raters 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 in MFRM. Results show that in both analyses, the percentage of observed exact agreements are close to the Rasch model's expectations. In FullA, the observed agreement (36.7%) is close to the Rasch model expectation (37%). Similarly, in IELTSA, observed agreement (35.6%) is close to the model expectation (37.9%). These findings suggest that the marking behaviour of the raters in the study aligns to the Rasch model expectations, that is, raters acting as independent experts.

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 … similar ratings on conceptually distinct traits' (Myford and Wolfe 2004:209) or in the words of Yorozuya and Oller Jr (1980:136) 'a kind of [judgement bias] spillover across scales causing them to be more strongly correlated with each other'.

Myford and Wolfe (2000:42) suggest a set of criteria that can be used for detecting different types of rater effects. Table 5.5 shows the type of rater effect in the first column and the infit and outfit limits associated with each effect in the second and third columns, respectively. 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.

We have already discussed rater consistency (accuracy) and underfit (random effect) with all four raters falling within stringent lower and upper infit and outfit control limits of 0.7 and 1.3.

Evidence for central tendency and/or halo effect is infit and outfit values below 0.7 (Myford and Wolfe 2000:42) or below 0.5 (Linacre 2018a:249), which would be considered as 'muted', indicating very little variation in score assignment across categories. In contrast, an infit index between 0.7 and 1.3 but with a corresponding outfit larger than 1.3 is considered to be 'noisy', indicating unexpected and inconsistent irregularities and evidence of extreme category overuse (Linacre 2018a, Myford and Wolfe 2000). Results of individual raters in the table suggest that the presence of these systematic rater effects is unlikely.

To summarise, the four raters in this study exercised severity levels which were significantly different from one another. Nevertheless, raters exhibited high levels of self-consistency in their marking 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 is therefore not considered problematic, as differences in severity are directly parameterised in the MFRM model and examinee ability estimates are adjusted accordingly (Linacre 2018a). An examination of rater fit indices did not reveal the influence of other systematic rater effects either. We can therefore conclude that the measurement error associated with systematic rater influences is largely controlled for in the study, lending support to the reliability of the adjusted examinee measures.

Table 5.5 Rater effect criteria and results

Rater effect	Infit criteria	Outfit criteria	Rater (R) ID	Infit results (IELTS scale)	Outfit results (IELTS scale)	Infit results (TD scale)	Outfit results (TD scale)
Accurate	0.7≤infit≤1.3	0.7≤infit≤1.3	R1	1.12	1.17	1.10	1.20
Random	infit >1.3	outfit >1.3	R2	0.96	0.94	1.05	1.03
Halo/Central	infit<0.7	outfit<0.7	R3	0.72	0.74	0.79	0.83
Extreme	0.7≤infit≤1.3	outfit>1.3	R4	1.12	1.13	1.05	1.03

Note: Rater effect criteria are based on Myford and Wolfe (2000:42).

The criterion facet

The difficulty of each criterion, similar to other facet elements, has been mapped onto the linear logit scale. The measurement report for the five analytic criteria used in the current study (FC, LR, GA, P, and TD) is provided in Table 5.6 in increasing order of difficulty.

Table 5.6 Rating scale criteria measurement report

Criterion	Observed average	Fair-M average	Measure	Model SE	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
SD	1.3	1.28	0.66	0.00	0.29	0.46

Model, Sample: RMSE .04, Adj (True) SD: .66, Separation 18.32, Strata: 24.76, Reliability: 1.00
Model, Fixed (all same) chi-square: 1350.5, *df*: 4, Significance (probability): .00

The IELTS analytic criteria (FC, LR, GA, P) have similar difficulty values and were located within close proximity of each other on the facet map (Figure 5.2), with FC as the easiest criterion ($\delta_{FC} = 0.14$) and P as the most difficult criterion ($\delta_P = 0.47$) covering a range of 0.33 logits. All infit and outfit statistics are within the acceptable quality control range of 0.5 to 1.5. Strikingly different in its logit value is the TD criterion ($\delta_{TD} = -1.15$), which is the easiest criterion for examinees to achieve a high score on. It is also the criterion with the highest associated infit value of 1.43. While falling within the acceptable quality control range, its comparatively higher infit value and markedly lower difficulty level suggest that the criterion is functioning in a distinct way from the other criteria.

The *trait separation indices* at the bottom of the table (G = 18.32, H = 24.76) and the reliability value of 1.00 suggest almost 25 statistically distinct difficulty strata. The null hypothesis that all criteria are of similar difficulty is therefore rejected ($\chi^2 = 1350.5$, $df = 4$, $p = 0.00 < 0.01$). Given the marked contrast in the difficulty levels of the IELTS criteria and the TD criterion, this finding is not surprising. I therefore independently examined the same statistics for the IELTS criteria to establish the extent to which the criteria are statistically distinct.

Measurement results for the IELTS criteria are presented in Table 5.7. The *trait separation indices* (G = 6.26, H = 8.69) and the separation reliability value of 0.98 suggest that criteria can be reliably separated into approximately eight difficulty levels. Moreover, the significant chi-squared value ($\chi^2 = 118.9$, $df = 3$, $p = 0.00 < 0.01$) confirms that the traits are statistically distinct and therefore not redundant. In other words, these results serve as counter-evidence for the presence of a halo effect in the IELTS rating scale (Myford and Wolfe 2004).

Table 5.7 IELTS criteria measurement report

Criterion	Observed average	Fair-M average	Measure	Model SE	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
SD	0.5	0.48	0.25	0.00	0.15	0.13

Model, Sample RMSE .04, Adj (True) SD: .24, Separation: 6.26, Strata: 8.69, Reliability: .98
Model, Fixed (all same) chi-square: 118.9, *df*: 3, Significance (probability): .00

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.

Rating scale functioning

The structure of the rating scales for each criterion used in the study can be visually inspected in the facet map (Figure 5.2). The IELTS criteria generally display a similar category structure and TD, is, as expected, different from the IELTS criteria, given the fewer number of score categories. FACETS produces a series of category statistics which can be used to evaluate scale functionality. In a later section I will use this same information for a fine-grained analysis of the effects of topic difficulty on scores across criteria and at different points on the scales.

Let us first focus on a single criterion of FC; the category statistics for each band level of the FC scale are presented in Table 5.8. The first column in the table shows the IELTS bands – those observed in the data – followed by the count and percentage of observations for each score. Note 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.

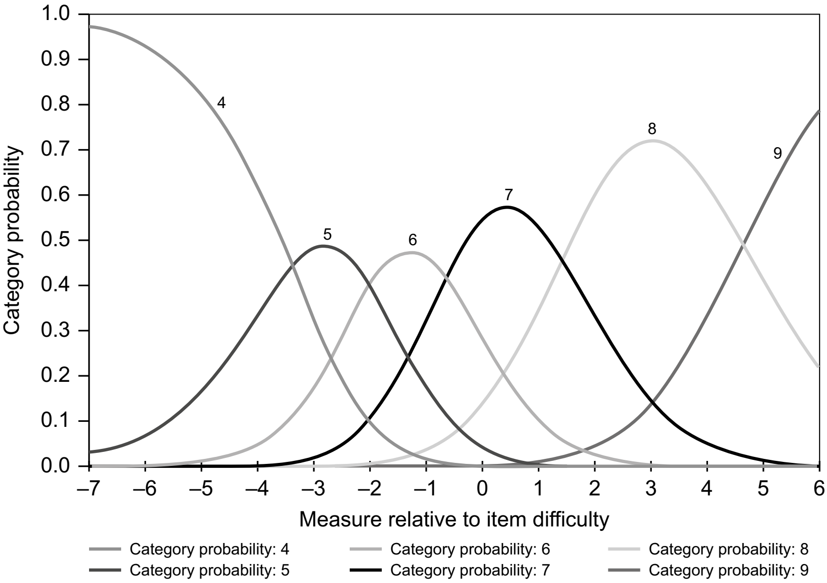
Table 5.8 Category statistics: Fluency and Coherence

Data			Quality control			Rasch-Andrich thresholds	
Score	**Used**	**Counts (%)**	**Average measure**	**Expected measure**	**Outfit MnSq**	**Measure**	**SE**
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 observed in each score category. The model 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 are also provided in the ‘expected measure’ column. Ideally, the observed and expected figures exhibit similar values; however, should the differences be large, the misfit would be captured in the ‘outfit mean square’ statistic column which, as a rule of thumb, should not exceed 2.00 (Eckes 2009).

The final columns – the Rasch-Andrich thresholds or step calibrations as used in FACETS – 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 correct scale functioning is visually depicted in Figure 5.3, where there is no evidence of disordered thresholds and as the probability of achieving a lower score (e.g. band score 4) decreases, the probability of the next adjacent score (band score 5) increases.

Figure 5.3 Category probability curves for Fluency and Coherence (FC)



The category statistics and probability curves for all remaining criteria were evaluated using a similar approach with results suggesting that the rating scales for different criteria were functioning appropriately in general[2].

Global model fit

With the exception of the topic facet (focus of next section), we have considered all important facets of our assessment context so far in the chapter. A final consideration is to examine 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 the 'practical utility' of the model that 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 the table of unexpected responses. 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

[2] The main exception was the P criterion, where a problem with disordered categories was observed. A closer examination suggested that it was the odd bands – where there are no specific descriptors and instead reference is made to adjacent bands – that was contributing to the problem. As this was an issue with the original scale, no further remedial action was taken.

the percentage of standardised residuals which fall outside of the criterion range of −3.00 to +3.00 are 1% or less (Linacre 2018a). Considering the present sample, the results are as follows, with both sets of results suggesting a satisfactory fit of the data to the Rasch model.

- FullA: Total number of valid responses = 7,010; standardised residuals ≥ |2| = 100 and standardised residuals ≥ |3| = 41 i.e. 1.43% ≥ |2| and 0.58% ≥ |3|.
- IELTSA: Total number of valid responses = 5,608; standardised residuals ≥ |2| = 100 and standardised residuals ≥ |3| = 23 i.e. 1.78% ≥ |2| and 0.41% ≥ |3|.

Focus on the topic facet

In the previous sections, we looked at the results for the main facets of analyses, which provided supportive evidence for the quality of the speaking data and the fit of the data to the Rasch model. In this section, we focus on the topic facet in order to answer some of the main RQs.

To remind the reader, the IST 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 labelled the *topic* facet for ease of referencing. A total of 18 topics are used in the research, distributed across task types as follows: Task Type A (six topics), Task Type B (four topics), and Task Type C (eight topics). Individual tasks/topics can be accessed in Appendix B.

In the main run of the MFRM analysis, topic is defined as an independent facet consisting of 18 elements regardless of task type. This allows for a general picture of the spread of topics and the sequencing of task types. Additional analyses were subsequently run for each task type in order to examine the spread of topic difficulties within each task type. Similar to previous sections, separate analyses were carried out for all criteria together (FullA) and for the IELTS criteria (IELTSA). A further analysis focusing specifically on the TD criterion was also carried out, as I hypothesised that topic is likely to have the largest influence on the TD scores.

Topic difficulty range

A visual inspection of the facet map in Figure 5.2 illustrates the narrow range of the topic difficulty distribution compared to the examinee ability distribution. Table 5.9 summarises these distributions for the three analyses (FullA, IELTSA, and TD only).

Results show 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 more pronounced once TD is removed (IELTSA); 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. A markedly different pattern is observed for the TD criterion analysis, where the observed examinee has dropped to 5.58 whereas the topic difficulty range has increased to 2.12 thus reducing the ratio to 2.63. The examinee separation indices (last column) 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 5.9 Examinee ability range compared to topic difficulty range

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 findings can help 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. On this basis, I argue that the TD criterion is functioning as intended in the study design 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 BK of topic, variability in TD scores is influenced by the underlying unidimensional speaking proficiency construct that the speaking tasks are designed to measure. This lends partial support to the use of a content-oriented criterion in speaking assessment contexts (in line with Sato 2012), a theme that I will return to 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. Let us now take a look at the topic measurement reports in more detail as presented in Table 5.10 and Table 5.11 for FullA and IELTSA, respectively. The first three columns provide the topic reference number, the task type, and a short topic description. Topics are ordered in ascending difficulty (logits). The remaining columns include statistical information similar to that discussed for previous facets.

Results show that when all criteria are included (FullA), the speaking tasks exhibit 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 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 ($\chi^2=170.3$, $df=17$, $p=0.00<0.01$) 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; 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. What needs to be further established, however, is whether the observed progression in task type difficulty matches the expected progression: Task Type A measures < Task Type B measures < Task Type C measures.

The second column in Table 5.10 does not reveal a clear or consistent pattern of progression along task types in increasing difficulty. Task Type B topics appear to be easier than Task Type A topics with the exception of Topic B.2, which is closer in difficulty to Task Type C topics. Some Task Type A topics (e.g. A.3 and A.4) exhibit difficulty measures similar to Task Type C topics, which is against expectations, as Task Type A topics are designed to be the easiest. Task Type C topics more 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 data.

Table 5.11 presents the topic measurement report but this time with the TD criterion removed from the analysis. Immediately noticeable are changes in the difficulty measures of topics as well as the rankings. Let us assign a ranking of 1 for the easiest topic and 18 for the most difficult topic. Topic A.5 ranked seventh in FullA but the ranking shifted to second in IELTSA. Similarly, Topic B.2 ranked twelfth in FullA but was subsequently ranked as the second most difficult topic (ranked seventeenth) in IELTSA. These

Table 5.10 Topic measurement report (FullA)

Topic ID	Task type	Topic	Observed average	Fair-M average	Measure	Model SE	Infit MnSq	Outfit MnSq
A.6	A	Dancing	5.9	5.62	−0.31	0.07	1.06	1.42
B.4	B	Describe an important choice	5.8	5.61	−0.29	0.08	1.12	1.14
B.1	B	Describe a friend	5.4	5.6	−0.27	0.07	0.99	1.12
B.3	B	Describe someone in your family	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	Describe a river, lake or sea	5.2	5.38	0.12	0.07	1.2	1.21
C.8	C	Choices in everyday life	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	The economic importance of rivers, lakes and the sea	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
SD			0.2	0.13	0.22	0.01	0.12	0.19

Model, Sample: RMSE .07, Adj (True) SD: .21, Separation: 2.97, Strata: 4.29, Reliability: .90
Model, Fixed (all same) chi-square: 170.3, *df*: 17, Significance (probability): .00

findings illustrate how the inclusion/exclusion of the TD criterion impact topic difficulty measures as well as task rankings.

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. 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 ($\chi^2 = 104.9$, $df = 17$, $p = 0.00 < 0.01$).

When examining the progression in difficulty levels of topics in terms of task type, Table 5.11 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 levels. Moreover, Task Type B topics are no longer clustering at the top of the table, that is, the location for the easiest topics. Rather, their location is shifted down while still interspersed with Task Type A topics. Contrary to expectations, Topics A.3 and A.4, which are designed to be the easiest task types (A), are closest in difficulty to Task Type C. I will draw on these results to answer the following RQs:

- *To what extent are the topics of speaking tasks used in parallel versions of a language proficiency interview similar in terms of difficulty?*
- *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 IST can be reliably divided into approximately three or four statistically distinct difficulty levels. This is as expected: the tests are designed to include three task types of increasing difficulty. The sequencing of topics, however, did not match the expected progression in task type difficulty.

The most consistent pattern was observed for Task Type C – designed as the most difficult task type – where the majority of topics exhibited the highest difficulty levels. The distinction between Task Type A and Task Type B topics, on the other hand, was not clear-cut and their respective topics did not cluster around specific 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. These results raise questions regarding the cognitive validity of the speaking tasks, as the tasks do not necessarily increase in difficulty in line with what test developers had in mind. Note that

Table 5.11 Topic Measurement Report (IELTSA)

Topic ID	Task type	Topic	Observed average	Fair-M average	Measure	Model SE	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	Describe a friend	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	Describe an important choice	6.3	6.1	−0.15	0.09	1.06	1.05
B.3	B	Describe someone in your family	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	Choices in everyday life	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	Describe a river, lake or sea	5.7	5.9	0.23	0.09	1.08	1.09
C.4	C	The economic importance of rivers, lakes and the sea	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
SD			0.2	0.10	0.20	0.01	0.10	0.11

Model, Sample: RMSE: .08, Adj (True) SD: .19, Separation: 2.21, Strata 3.27, Reliability: .83
Model, Fixed (all same) chi-square: 104.9, *df*: 17, Significance (probability): .00

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.

Topic effects at the test level

In order to evaluate the topic effect from a score perspective, we can draw on information from the topic measurement reports and the category statistics. This allows for a close examination of differences in topic difficulty in relation to average examinee measures at different score categories across criteria both at the test and task levels, and for addressing some of the study's RQs. Once again, all results are reported separately for FullA and IELTSA.

In the operational IST, an overall band score is awarded at the end of the test for all speaking tasks covered in the three test parts. In my study, on the other hand, raters awarded marks for each speaking task so that each topic would have an estimated difficulty measure. Given that each form of the IST is comprised of two Task Type A topics, one Task Type B topic, and two Task Type C topics, then a combination of topics (from each task type) can be used to create multiple parallel forms. For illustration purposes and to evaluate the effects of topics at the test level, I decided to construct two parallel forms from available topics: one containing the easiest topics (within each task type) and one containing the most difficult topics (within each task type). The average difference between the easiest and most difficult forms would thus represent *the maximum* difference between two tests owing to topic difficulty. This difference can then be examined in relation to average examinee measures at different score categories across criteria to provide a fine-grained analysis of topic effects on scores at the test level.

I used the topic difficulty measures from Table 5.10 (FullA) to construct the two 'easy' and 'difficult' parallel forms from the combination of topics within each task type (see Table 5.12). The average difficulty measures for the 'easy' and 'difficult' forms are −0.17 logits and 0.18 logits, respectively, which brings the difference between them to 0.35 logits. At the test level, this is the difference that can be attributed to the selected topics.

Let us now contextualise the value of 0.35 logits in relation to the average examinee ability necessary to move across adjacent band levels on different criteria. This will allow us to infer whether the 0.35 logit difference can have any meaningful (practical) implications in terms of the rating scales. I will illustrate this with an example; the category statistics for FC (Table 5.8) show that the average examinee ability measure at Band 4 is −2.92 and the average measure at Band 5 is −1.98. An increase in ability of **0.94 logits** is therefore necessary to move from Band 4 to Band 5 on the FC scale. The difference of 0.35 logits – attributed to topic difficulty across the easy and difficult forms – is not sufficient for moving across these two adjacent bands. The same approach can be applied to the remaining bands and criteria to evaluate the topic effect at test level.

The increase in average ability measures for all adjacent band levels and across criteria was subsequently calculated and summarised in Table 5.13. Results show that there are no score categories where the average ability measure required for moving across adjacent bands exceeds the logit value of 0.35 – attributed to differences in topic difficulty measures 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 0.35 between the two test forms (0.89 > 0.35). This is also the case for the easiest criterion (TD); the smallest average ability measure required to move across two adjacent bands is 0.44 for Levels 3–4, which once again exceeds 0.35 (0.44 > 0.36). These results show that even when the two speaking test forms include the easiest versus the most difficult topics, examinee ability measures are unlikely to be influenced by differences in topic difficulties.

I repeated the above analyses for the IELTS criteria (IELTSA) by first constructing two test forms consisting of the easiest and most difficult topic combinations, calculating the average differences, and comparing this value against the average ability measures necessary to move along adjacent band score categories across the IELTS criteria. In reporting the results, I will draw on the Fair-M average results to help further contextualise the findings in terms of the original IELTS scale.

The difference in average topic measures between the ‘easy’ and ‘difficult’ parallel forms is 0.30 logits (Table 5.14). In reference to the original IELTS scale, the Fair-M average values for the ‘easy’ and ‘difficult’ forms are 6.11 and 5.95, respectively – a difference of 0.16 IELTS band scores. This difference is smaller than the smallest meaningful unit in IELTS which is 0.5 band scores (0.16 < 0.5). It is safe to argue that any differences in topic difficulty are unlikely to have a meaningful (practical) impact on candidate scores at the test level.

When focusing on different band levels across criteria (Table 5.15), a similar pattern emerges. On average, an increase in ability of 1.42 logits is necessary to move across adjacent band scores across the different IELTS criteria. The difference of 0.30 logits in topic difficulty measures is thus too small in comparison to the speaking ability required for examinees to move to a higher band. Put differently, test takers’ performance measures at the test level are unlikely to be influenced by the difficulty level of topics.

I would also like to draw attention to another observation regarding the ‘average’ ability measures required to move across adjacent bands in the FullA vs. the IELTSA; in the former, the logit measure hovers around 1.0 whereas in the latter, the logit measure hovers around 1.4. This finding implies that when the TD criterion is included, it is easier for examinees to move across band levels. In other words, the TD criterion may have a facilitative

Table 5.12 Construction of ‘easy’ and ‘difficult’ parallel forms of IST (FullA)

Parallel form (easy)					Parallel form (difficult)				
Topic ID	Task type	Topic	Fair-M average	Measure	Topic ID	Task type	Topic	Fair-M average	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	Describe an important choice	5.61	−0.29	B.2	B	Describe a river, lake or sea	5.38	0.12
C.5	C	Family similarities	5.48	−0.06	C.4	C	Economic importance of rivers, lakes and the sea	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) = 0.35 (logits)									

Table 5.13 Increase in average ability measures in adjacent categories for all criteria (FullA)

Bands	FC	LR	GA	P[3]	TD (levels)	
Bands 3–4				0.67		
Bands 4–5	0.94	0.99	1.07	0.84		
Bands 5–6	0.90	0.91	0.93	0.74	Levels 1–2	0.99
Bands 6–7	0.89	0.90	0.90	1.82	Levels 2–3	0.97
Bands 7–8	1.67	1.5	1.55	0.99	Levels 3–4	0.44
Bands 8–9	1.25	1.29	1.09	0.74	Levels 4–5	0.93
Average	1.13	1.12	1.11	1.03		0.83

effect on achieving higher scores. The findings from the above analyses can be drawn upon to answer the following RQ: *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?*

Through a series of fine-grained analyses, I have demonstrated that there are minimal 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 – whether at the 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, we can infer 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.

Topic effects at the task level

In this section, we will look at topic effects at the task level and examine the topic measurement reports for each task type. Similar analytic procedures from the previous section are repeated here but this time we will consider the differences between the easiest and most difficult topics *within each task type*.

Task Type A

The topic measurement reports for Task Type A topics are presented in Table 5.16 and Table 5.17 for FullA and IELTSA, respectively. All topic infit and outfit statistics fall within the stringent range of 0.7 to 1.3. The ranking of topics in ascending order of difficulty is similar across analyses with Topic A.6 (Dancing) as the easiest and Topic A.3 (Festivals) as the most difficult.

[3] Due to the problems with reversed thresholds for the P criterion, the observed average value was replaced with the expected value for the disordered categories.

Table 5.14 Construction of 'easy' and 'difficult' parallel forms of IST (IELTSA)

Parallel form (Easy)					Parallel form (difficult)				
Topic ID	**Task type**	**Topic**	**Fair-M average**	**Measure**	**Topic ID**	**Task type**	**Topic**	**Fair-M average**	**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	Describe a friend	6.14	−0.23	B.2	B	Describe a river, lake or sea	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	Economic importance of rivers, lakes and the sea	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) = 0.30 (logits)
Differences in Fair-M average measures = 6.11−5.95 = 0.16 IELTS band scores

Table 5.15 Increase in average ability measures in adjacent categories for all criteria (IELTSA)

Bands	FC	LR	GA	P[4]
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

Table 5.16 Measurement report for Task Type A topics (FullA)

Topic	Observed average	Fair-M average	Measure	Model SE	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
SD	0.2	0.07	0.13	0.01	0.08	0.07

Model, Sample: RMSE .07, Adj (True) SD: .11, Separation: 1.63, Strata 2.51, Reliability: .73
Model, Fixed (all same) chi-square: 17.8, *df*: 5, Significance (probability): .00

The spread of difficulty measures, however, is wider in FullA, spanning 0.3 logits, whereas this range is limited to 0.10 logits in IELTSA.

To determine whether the observed range in topic measures is associated with statistically distinct difficulty levels, I examined the separation statistics. In FullA, the topic separation indices (G = 1.63, H = 2.51, r = 0.73) suggest that the six Task Type 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 r = 0.73. 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 Type A topics cannot be considered parallel versions of the same task.

The topic separation indices in IELTSA, in contrast, portray a different picture. Once the TD criterion is removed from the analysis, a drop is

[4] Due to the problems with reversed thresholds for the P criterion, the observed average value was replaced with the expected value for the disordered categories.

Table 5.17 Measurement report for Task Type A topics (IELTSA)

Topic	Observed average	Fair-M average	Measure	Model SE	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
SD	0.2	0.05	0.1	0.01	0.13	0.13

Model, Sample: RMSE .08, Adj (True) SD: .06, Separation .76, Strata: 1.34, Reliability: .36
Model, Fixed (all same) chi-square: 7.7, *df*: 5, Significance (probability): .17

observed in the separation, strata, and reliability values (G = 0.76, Strata = 1.34, r = 0.36), indicating that the six topics cannot be separated into statistically distinct difficulty levels and can thus be considered parallel; a result also substantiated by the non-significant chi-squared statistic.

Task Type B

The same analyses were repeated for the four Task Type B topics (see measurement reports for FullA and IELTSA in Table 5.18 and Table 5.19). All topic infit and outfit statistics fall within the stringent range of 0.7 to 1.3.

Topic rankings vary across the two analyses, with the exception of Topic B.2 (Describe a river, lake or sea), which remains the most difficult Task Type B topic. 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 difficulty range in Task Type A topics.

Table 5.18 Measurement report for Task Type B topics (FullA)

Topic	Observed average	Fair-M average	Measure	Model SE	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	1.00	1.11
SD	0.2	0.11	0.19	0.01	0.11	0.16

Model, Sample: RMSE .08, Adj (True) SD: .23, Separation: 2.93, Strata: 4.24, Reliability: .90
Model, Fixed (all same) chi-square: 26.8, *df*: 3, Significance (probability): .00

Table 5.19 Measurement report for Task Type B topics (IELTSA)

Topic	Observed average	Fair-M average	Measure	Model SE	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
SD	0.2	0.13	0.25	0.02	0.09	0.1

Model, Sample: RMSE .07, Adj (True) SD: .18, Separation: 2.73, Strata: 3.97, Reliability: .88
Model, Fixed (all same) chi-square: 21.8, *df*: 3, Significance (probability): .00

Unlike Task Type A topics, the results of the two analyses for Task Type B topics exhibit more similarity in terms of high topic separation indices and associated reliability values ($G_{FullA} = 2.93$, $H_{FullA} = 4.24$, $r_{FullA} = 0.90$; $G_{IELTSA} = 2.73$, $H_{IELTSA} = 3.97$, $r_{IELTSA} = 0.88$). These results suggest that Task Type B topics are not equivalent in difficulty (with or without the TD criterion) and that they can be separated into approximately four statistically distinct difficulty strata. The significant chi-squared results for both analyses confirm that Task Type B topics do not share the same difficulty measures.

A closer examination of the relative topic difficulty measures illustrates that it is only Topic B.2 which has a markedly different difficulty measure compared to the other three Task Type B topics. A hypothesis is that this topic is qualitatively different from the other topics. This task type requires examinees to describe ‘a friend’ (B.1), ‘a river, lake or a sea that you like’ (B.2), ‘someone in your family’ (B.3), and ‘an important choice you had to make’ (B.4). The common theme amongst Topics B.1, B.3, and B.4 is that they appear to be more personal than the ‘river’ (B.2) topic. A possible explanation for the findings is that the test takers were better able to relate to the more personal topics.

A final consideration before moving to Task Type C topics is whether the statistically distinct difficulty strata for Task Type B topics can have a practical effect on performance scores in the IELTS context. To address this question, I examined the Fair-M average results for the topics in IELTSA. Topic B.4 has the highest Fair-M average mark of 6.16 whereas Topic B.2 has the lowest Fair-M average mark of 5.89. The difference of 0.27 is less than half a band, indicating that the statistically significant difference in the topic difficulty measures does not translate into a practical influence on performance scores at the task level across criteria.

Task Type C

The topic measurement reports for the eight Task Type C topics are reproduced in Table 5.20 and Table 5.21 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 with variations in the ranking of the topics in between.

Table 5.20 Measurement report for Task Type C topics (FullA)

Topic	Observed average	Fair-M average	Measure	Model SE	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
SD	0.2	0.12	0.23	0.01	0.12	0.2

Model, Sample: RMSE .07, Adj (True) SD: .22, Separation: 3.05, Strata: 4.39, Reliability: .90
Model, Fixed (all same) chi-square: 106.2, *df*: 7, Significance (probability): .00

Table 5.21 Measurement report for Task Type C topics (IELTSA)

Topic	Observed average	Fair-M average	Measure	Model SE	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
SD	0.2	0.1	0.21	0.01	0.1	0.12

Model, Sample: RMSE .08, Adj (True) SD: .19, Separation: 2.29, Strata: 3.38, Reliability: .84
Model, Fixed (all same) chi-square: 64.2, *df*: 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 three to four statistically distinct difficulty levels ($G_{\mathrm{FullA}} = 3.05$, $H_{\mathrm{FullA}} = 4.39$, $r_{\mathrm{FullA}} = 0.90$; $G_{\mathrm{IELTSA}} = 2.29$, $H_{\mathrm{IELTSA}} = 3.38$, $r_{\mathrm{IELTSA}} = 0.84$). The significant chi-squared results substantiate the finding that 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. This value is smaller than the smallest unit which can have a meaningful difference in IELTS. It is therefore unlikely for the differences in difficulty of Task Type C topics to have a meaningful and practical effect on performance scores at the task level ($0.2 < 0.5$).

So far, the analyses have focused on topic effects at the task level across criteria. What remains to be seen is whether any differences in topic measures have an influence at specific score categories at the criterion level.

Table 5.22 summarises the information extracted from category statistics for each criterion and for the different task types from FullA. The differences between average ability measures observed at adjacent band scores across criteria were separately calculated. Also included in the table is the maximum difference between the easiest and most difficult topic measures within each task type, extracted and calculated from the topic measurement reports.

Table 5.22 Increase in average ability measures in adjacent categories (FullA)

	Bands	FC	LR	GA	P	TD (Levels)	
Task Type A	Bands 3–4				1.21		
	Bands 4–5	1.20	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						
	Bands	**FC**	**LR**	**GA**	**P**	**TD (Levels)**	
Task Type B	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	1.40	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						

Table 5.22 (continued)

	Bands	**FC**	**LR**	**GA**	**P**	**TD (Levels)**	
Task Type C	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						

For Task Type A, the maximum difference between the easiest and most difficult topic is 0.30 logits. To see whether this value can have a practical (meaningful) impact on performance, we examine the value in relation to the average ability levels required to move across different score categories and for the different criteria. 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 topics. The only instance where the 0.30 value might have an impact is for the TD criterion; the table shows that an increase of 0.32 logits is required to move from Level 3 to Level 4. Given that 0.30 is quite close to 0.32, we can argue 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 in Task Types B and C topics 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.

The above analyses were repeated for IELTSA with results summarised in Table 5.23. The maximum differences between the easiest and most difficult topics are 0.24, 0.52, and 0.54 for Task Types A, B, and C topics, respectively. In terms of the IELTS raw-score metric, these values correspond to 0.12, 0.27, and 0.27 IELTS bands, respectively, all of which fall below the minimum meaningful unit of 0.5 band scores in the IELTS scale. Not only have the

Table 5.23 Increase in average ability measures in adjacent categories (IELTSA)

	Bands	**FC**	**LR**	**GA**	**P**
Task Type A	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				
	Bands	**FC**	**LR**	**GA**	**P**
Task Type B	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				
	Bands	**FC**	**LR**	**GA**	**P**
Task Type C	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				

maximum differences in topic difficulty levels decreased in comparison with the previous analysis (FullA), the average ability levels required to move along adjacent score categories across the different criteria have also increased to approximately 1.5 logits, which render the likelihood of a topic influence on scores minimal. Put differently, 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 to address the following RQs: *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?*

Findings have shown 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 statistically distinct difficulty strata. In other words, there are at least two topics within each task type that could not be considered parallel. Comparable results emerged for Task Types B and C topics in IELTSA. The six Task Type A topics, however, exhibited very similar difficulty measures and could therefore be considered parallel.

In terms of practical significance, results of the analyses (FullA and IELTSA) suggest that differences in topic difficulty measures (within task types) are not large enough 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, the differences were not large enough to translate into meaningful differences in performance scores. The only exception where an easier or more difficult topic can potentially result in examinees achieving a higher or lower score is for the TD criterion but the effect is only likely to be limited to specific levels.

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. These two sets of findings are reported in succession, as they can provide complementary perspectives on the effects of topic on performance at the task level.

Topic effects: A language functions perspective

Another way of looking at topic effects is to consider their influence from the perspective of functions elicited. In the IST, each task type is intended to elicit specific functions from test takers. We can therefore argue that when task type is held constant, different (parallel) topics should elicit a similar range of functions. This was empirically tested in the research and is the focus of this section.

To remind the reader, raters in the study were asked to complete an ‘observation checklist’: an instrument that allows for a comparison of different speaking tasks in terms of the range of functions they can elicit (see Chapter 4 for more details). The data from the observation checklist was sorted by task type so that topics within each task type could be compared in terms of the range of elicited functions. Note that raters identified *types* of functions observed in each performance and not the *frequency* with which each function was observed. Raters reached an exact agreement of 79.2%, on average, for the common batch performances.

In the interest of space, I will not report on the full results but use illustrative examples instead. Table 5.24 lists the different types of functions in the observation checklist in the first column. In the remaining columns, two topics from each task type were selected: an easy topic denoted by the (+) sign and a difficult topic denoted by the (−) sign on the basis of topic difficulty measures. A description of each topic is provided at the bottom of the table. The data in the table is expressed in percentages, that is, 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 'providing 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 Type A topics are designed to be 'familiar' topics, the small percentage of observations for the personal information function is likely to 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 measure closer to Task Type C topics (designed to be more abstract/unfamiliar).

Table 5.24 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
Analysing	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

Table 5.24 (continued)

Task type	**A**		**B**		**C**	
Function	**A.6 (+)**	**A.3 (−)**	**B.1 (+)**	**B.2 (−)**	**C.5 (+)**	**C.6 (−)**
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, lake or sea; 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. When all Task Type A topics were taken into consideration, the speculating function was observed about 5–11% of the time with the exception of A.6. In terms of the other functions, the percentages appear to be more similar.

If we now consider the two Task Type B topics – Topic B.1 (Describe a friend) and Topic B.2 (Describe a river, lake or sea) – and focus on the proportion of personal to non-personal functions, we see a sharp drop in the percentage of the 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 percentages of personal information functions whereas more difficult topics are associated with lower percentages of personal information functions. We can therefore indirectly infer that topics that have been shown to be psychometrically easier require test takers to draw more extensively on personal information likely to be more readily available to them.

Other noticeable differences between the percentage of observed functions in Task Type B topics are for functions such as 'expressing' and 'justifying opinions' and most markedly in 'expressing preferences', 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 for Topic B.1 although it was observed for Topic B.2 about 7% of the time, lending confirmatory evidence for the MFRM analyses where B.2 was found to exhibit a higher difficulty measure compared to B.1.

Finally, if we examine the percentages of observed functions in Topics C.5 (Family similarities) and C.6 (Genetic research), we see 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 markedly observed for Topic C.5 (30.8%) compared to C.6 (2.6%). An examination of the percentage of the occurrence of this function across topics and across task types reveals that the 'disagree' function is rarely observed and when it is observed, its percentage is limited to between 2 and 8%. Topic C.5 is the only exception with a substantively higher percentage of observations. In terms of the 'commenting 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.

These findings can be used to address the following RQ: *When task type is held constant, to what extent are the observed functions elicited by different topics similar?*

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. The examples provided serve to illustrate how different topics – with similar difficulty measures – may be qualitatively different in terms of the functions they elicit. Put differently, topics may not have an influence on performance scores but can nevertheless have an influence on the range of functions elicited within a given task type.

Before moving on, I would also like to draw attention to Table 5.25, which shows the average percentage of observations for each function across all speaking tasks, arranged in descending order. The data in the table clearly reflects the information-oriented nature of the IST as evidenced in the comparatively high frequency of the *informational functions* (e.g. expressing opinions, providing personal and non-personal information, justifying opinions, comparing, and contrasting) compared to the low frequency of observations for *interactional functions* such as negotiation of meaning, agreeing, disagreeing, and conversational repair, with the latter occurring only 0.2% in the data.

These findings bring into question the extent to which the speaking test is successful in eliciting interaction. Moreover, they highlight the central role of *information* in these tests as the basis around which performance is built, lending support to the potential influence of BK.

Role of background knowledge

In presenting the conceptual-psychometric framework (Eckes 2009) adopted for the study, we looked at how some test taker characteristics such as L1 or gender are typically considered as 'distal' factors and examined in interaction

Table 5.25 Average percentage of observed functions across tasks

Function	Average (%)
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
Analysing	4.7
Disagreeing	2.7
Conversation repair	0.2

or bias analyses rather than as ‘proximal’ facets. What I have argued, however, is that BK of topic is not a constant test taker characteristic, as it may vary from one topic to the next. As such, and given the research focus on evaluating the main effect of BK on performance, BK was explicitly parameterised as a proximal facet of assessment and a new MFRM analysis was defined as follows:

- examinee facet (81 participant elements)
- rater facet (four rater elements)
- criterion facet (five criteria elements)
- topic facet (18 topic elements)
- BK of topic facet (three condition elements).

In Chapter 4, I explained how BK measures for each person × topic combination were derived from the analysis of the BK questionnaires. These measures were subsequently divided into three groups: low, medium and high, constituting the three elements of the BK facet. The reason why BK measures could not be directly used in the analyses is because FACETS only accepts integer numbers. Linacre (2018a:247) suggests 'chunking this [continuous data] up into qualitatively advancing pieces and number the chunks.'

The BK facet: Effect on scores

The review of the literature suggested that higher levels of BK may have a facilitative effect on performance. Should this be the case, we can expect the low BK condition to be the most challenging condition for test takers and the high BK condition the least challenging and easiest. 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.

Results of the five-facet MFRM analysis are visually displayed in the vertical maps in Figure 5.4 and Figure 5.5 for FullA and IELTSA, respectively. The results of the BK facet now appear in the fifth column.

A glimpse of 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 thus warranting further analysis.

The BK measurement reports for FullA and IELTSA are presented in Table 5.26 and Table 5.27. The BK estimates for each condition are arranged in ascending order of difficulty (in logits) and illustrate that the high BK condition is associated with the lowest measures (high $BK_{FullA} = -0.32$, high $BK_{IELTSA} = -0.29$) and therefore easiest whereas the low BK condition has the highest measures (low $BK_{FullA} = +0.36$, low $BK_{IELTSA} = +0.34$), with difficulty measures spanning a range of 0.68 and 0.63 across the two analyses. The medium BK estimate remains the same in FullA and IELTSA. The infit mean square statistics for the three elements fall within the acceptable control limits of 0.5 to 1.5. In fact, the values range between 0.90 and 1.1 and are therefore very close to their expected value of 1.0.

The difficulty span suggests that the BK conditions are distinct 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

Figure 5.4 Facet map (FullA)

```
+----------------------------------------------------------------------------------+
|Ability (High)    |Severe |         Difficult        | FC  | LR  | GA  | P   | 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  |     | --- |
|     | ***       |       |       |          | P     |  7  |  7  |     |  6  |     |
|     | ***       | R2    | C.8   | LowBK    | FC    |     |     |     |     |     |
:     :           : R4    : C.1   :          : GA    :     :     :     :     :     :
:     :           :       : C.2   :          : LR    :     :     :     :     :     :
:     :           :       : B.2   :          :       :     :     :     :     :     :
:     :           :       : C.4   :          :       :     :     :     :     :     :
*   0 * ********* * R1    * 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 + ********  +       +       +          +       +     +     +     +     +     |
|     | ******    |       |       |          | TD    |  6  |  6  |  6  |     |     |
|     | ******    |       |       |          |       |     |     |     |  5  |     |
|     | **        |       |       |          |       |     |     |     |     |  2  |
|  -2 + **        +       +       +          +       + --- + --- + --- +     +     |
|     | *****     |       |       |          |       |     |     |     |     |     |
|     | *         |       |       |          |       |     |     |     | --- |     |
|     | *         |       |       |          |       |  5  |     |     |     | --- |
|  -3 + **        +       +       +          +       +     +  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                                |
|SD   |  1.37      | 0.47  |  0.16  | 0.34     | 0.67                                |
+----------------------------------------------------------------------------------+
```

Note: Each star (*) in the second column represents one examinee.
Measr = Measure, Cri = Criteria, FC = Fluency and Coherence, LR = Lexical Resource, GA = Grammatical Range and Accuracy, P = Pronunciation, TD = Topic Development.

Figure 5.5 Facet map (IELTSA)

```
+---------------------------------------------------------------------------+
|Ability (High)  |Severe |         Difficult        | FC  | LR  | GA  | P   |
|Measr|+examinee |-Rater |-Topic |-BKGroup  |-Cri   | S.1 | S.2 | S.3 | S.4 |
|-----+----------+-------+-------+----------+-------+-----+-----+-----+-----|
|   5 +          +       +       +          +       + (9) + (9) + (9) + (9) |
|     |          |       |       |          |       |     |     |     |     |
|     |          |       |       |          |       |     |     |     |     |
|     |          |       |       |          |       |     |     |     |  8  |
|   4 + **       +       +       +          +       +     +     +     +     |
|     |          |       |       |          |       |  8  |  8  |  8  |     |
|     | *        |       |       |          |       |     |     |     |     |
|     |          |       |       |          |       |     |     |     | --- |
|   3 +          +       +       +          +       +     +     +     +     |
|     | *        |       |       |          |       |     |     |     |     |
|     |          |       |       |          |       |     |     |     |  7  |
|     | *        |       |       |          |       |     |     |     |     |
|   2 + **       +       +       +          +       +     +     + --- +     |
|     |          |       |       |          |       | --- | --- |     | --- |
|     | *        |       |       |          |       |     |     |     |     |
|     |          |       |       |          |       |     |     |     |     |
|   1 +          +       +       +          +       +     +     +     +     |
|     |          |       |       |          |       |     |     |  7  |     |
|     | *        | R2    |       |          |       |  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  |     |     |     |
|  -2 + *****    +       +       +          +       +     +     +     +  5  |
|     | ****     |       |       |          |       |     |     |     |     |
|     | ****     |       |       |          |       | --- | --- | --- |     |
|     | *        |       |       |          |       |     |     |     |     |
|  -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                          |
|SD   |  1.73    | 0.73  |  0.18 | 0.32     | 0.25                          |
+---------------------------------------------------------------------------+
```

Note: Each star (*) in the second column represents one examinee.
Measr = Measure, Cri = Criteria, FC = Fluency and Coherence, LR = Lexical Resource, GA = Grammatical Range and Accuracy, P = Pronunciation.

Table 5.26 The BK measurement report (FullA)

BK level	Observed average	Fair-M average	Measure	Model SE	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
SD	0.2	0.2	0.34	0.00	0.09	0.15

Model, Sample: RMSE .03, Adj (True) SD: .34, Separation 12.13, Strata 16.51, Reliability .99
Model, Fixed (all same) chi-square: 285.2, *df*: 2, Significance (probability): .00

Table 5.27 The BK measurement report (IELTSA)

BK Level	Observed average	Fair-M average	Measure	Model SE	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
SD	0.2	0.16	0.32	0	0.09	0.1

Model, Sample: RMSE .03, Adj (True) SD = .32, Separation = 9.33, Strata = 12.78, Reliability = .99
Model, Fixed (all same) chi-square: 169.6, *df*: 2, Significance (probability): .00

assessment context, we also do not want BK conditions (an arguably construct-irrelevant factor) to exert a significant influence on performance. It is therefore desirable for the BK separation indices to be low and the separation reliability value to be close to 0. The results of the tables, however, indicate otherwise.

The separation indices for FullA ($G_{FullA} = 12.13$, $H_{FullA} = 16.51$, $r_{FullA} = 0.99$) 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 ($G_{IELTSA} = 9.33$, $H_{IELTSA} = 12.78$, $r_{IELTSA} = 0.99$) 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 ($\chi^2_{FullA} = 285.2$, $\chi^2_{IELTSA} = 169.6$) and the corresponding probability values of $p = 0.00 < 0.01$. These results confirm that the BK condition can have a statistically significant impact on performance.

What remains to be addressed is whether this statistically significant influence translates into practical significance in terms of performance scores.

The Fair-M average column in IELTSA shows that the difference between the high BK and low BK conditions is 0.3 IELTS band scores which, as discussed earlier, falls beneath the minimum IELTS band unit (0.5) which can have a practical and meaningful effect on candidate performance scores (0.3 < 0.5).

We can also exercise the same fine-grained approach used earlier to look at topic influence on scores by considering the effects of BK in relation to average examinee ability levels necessary to move across score categories in the different criteria. The average measures are calculated from the category statistics and summarised in Table 5.28 and Table 5.29 for FullA and IELTSA, respectively.

Table 5.28 Increase in average ability measures in adjacent categories (FullA)

Bands	FC	LR	GA	P	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 5.29 Increase in average ability measures in adjacent categories (IELTSA)

Bands	FC	LR	GA	P
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 (FullA) and 0.63 (IELTSA). 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 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 RQs: *Will differences in test takers' levels of BK 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 observed differences in levels of BK are statistically significant and can therefore pose distinct levels of challenge for test takers. This statistical significance, however, 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 specific band levels.

The above analyses have shown that, at least from a modelling perspective, BK can be explicitly parameterised as an additional facet in MFRM. A question that we 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. I would like to argue 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 BK 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 penalising 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.

The BK facet: Influence on other facets

In addition to examining the influence of BK on performance scores, its impact on other facets was also considered to see whether there are any marked changes in measurement results. To this end Table 5.30 and Table 5.31 provide a side-by-side comparison of the rater, topic, criterion, and examinee measurement results with and without the BK facet for FullA and IELTSA.

The rater facet results show the exact same ranking of raters with minimal differences (0.01–0.04 logits) in rater severity estimates. Negligible differences (0.01–0.02) are also observed for the criteria estimates. This means that the elements for these two facets remain 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, SD and ability range for each analysis are presented. In both FullA and IELTSA, the mean and SD of examinee ability have 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 (without BK) to 8.62 (with BK) is observed for IELTSA.

Table 5.30 MFRM measurement result comparisons (FullA)

	MFRM without BK			MFRM with BK		
	Facet elements	**Measure**	**Model SE**	**Facet elements**	**Measure**	**Model SE**
Rater	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 (Mean)	**Examinee (SD)**	**Ability range**	**Examinee (Mean)**	**Examinee (SD)**	**Ability range**
Examinee	−0.63	1.36	6.81	−0.63	1.37	7.08

When the differences between examinee measurement estimates with and without BK were calculated and examined for individual examinees, the maximum absolute difference was 0.35 logits for FullA and 0.32 logits for IELTSA (approximately 0.2 IELTS band scores). These differences are therefore not large enough to be translated into meaningful differences on scores (0.2 < 0.5 IELTS band).

Table 5.31 MFRM measurement result comparisons (IELTSA)

	MFRM without BK			MFRM with BK		
	Facet elements	**Measure**	**Model SE**	**Facet elements**	**Measure**	**Model SE**
Rater	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 (Mean)	**Examinee (SD)**	**Ability range**	**Examinee (Mean)**	**Examinee (SD)**	**Ability range**
Examinee	−1.18	1.72	8.49	−1.18	1.73	8.62

An examination of the topic facet reveals a strikingly different pattern compared to the other facets. Firstly, the topic rankings have changed in the analyses with/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). Secondly, 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 Topic C.6 (Genetic research) and Topic A.3 (Festivals) which exhibit a drop of 0.28 and 0.15 logits, respectively, in their difficulty measures. These are the topics for which examinees, on average, had reported the lowest BK. 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 the test taker characteristic of BK of topics. These findings are used to answer the following RQ: *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, BK was shown to have 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 varied when BK was explicitly parameterised in the analyses. This finding has important implications for 'objective' estimates of prompt/task/topic difficulty. The study has shown that difficulty of topics can be influenced by the BK that test takers bring to the test and therefore is not necessarily inherent to the task.

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. As explained in the methodology chapter, the linked design of the study was such that not all participants responded to all topics. This was in light of practical considerations and the fact that the Rasch model is robust against missing data (Eckes 2009) subject to sufficient linking in the data. The model looks at observed patterns in the data and on the basis of available information calculates estimations for unobserved (missing) data as well as 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 if the examinee did not respond to that particular task during data collection. The same is true for rater severity. However, a serious problem arises for the BK facet. The participants in the study completed the BK questionnaire only for those topics they attempted in the ISTs, therefore resulting in missing BK data. The Rasch model, in response, will look for patterns in BK groupings in order to make estimations for the missing BK 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, ironically, is the very one this research is trying to dismantle, given the argument that BK is highly individual and test taker dependent. Nevertheless, the assumption was necessary for running the analyses for the whole sample. To address this limitation, however, I used a different statistical approach where only the observed BK data was included thus circumventing the need to make any assumptions about levels of BK.

BK as predictor variable

The previous MFRM analysis had two limitations: firstly, the BK measures from the questionnaire could not be directly used in the analyses and had to be grouped into levels; and secondly, there was missing BK data for which the model made adjustments that were not necessarily justified in light of the individual nature of BK. To address these issues, I used a different statistical technique – multiple regression – to examine the extent to which BK as a variable 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 that have used more traditional statistical techniques.

In Chapter 4, I described how for every person × topic combination, an independent BK measure was calculated. To examine the influence of BK on a given topic, it was necessary to run another analysis to estimate a speaking ability measure for every person × topic combination. So far, all examinee speaking ability measures were based on the results of examinees responding to all 10 topics.

To estimate examinee measures on the basis of individual topics, I first rearranged the data so that each person × topic combination was treated as a distinct person and subsequently ran a three-facet MFRM analysis with examinee, rater, and criteria as facets. In this approach, topic is no longer considered a separate facet; instead, relative topic difficulties are absorbed in the resulting person × topic measures. An example (for person Z – ID 1) is provided in Table 5.32 for illustration purposes. In the first column, we see person Z responding to five different topics (1, 2, 9, 10, and 11). The New ID column shows that every person × topic combination (e.g. Person Z × Topic 1, Person Z × Topic 2) is identified with a different (new) reference. However, when the same person × 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 × topic measures are adjusted for relative rater severity. The 'BK grouping' 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 four-facet MFRM analysis (examinee, rater, topic, criteria) where the estimates remain the same for Person Z. In contrast, the 'topic ability measure' column shows the results from a three-facet MFRM analysis with examinee (person × topic), rater, and criteria as facets. This column illustrates variations in the ability measures of Person Z on the basis of the topic(s) they were assigned.

The approach for rearranging the data into a 'racked' data set and a three-facet MFRM analysis was repeated for all examinees (with and without the TD criterion). The resulting person × topic speaking measures

Table 5.32 Example of a racked data set for Person Z

Original ID	Topic	Rater	New ID	BK grouping	BK measure	Ability measure (10 topics)	Topic ability measure
1	1	3	1	High BK	4.72	2.52	2.73
1	2	4	2	High BK	5.45	2.52	2.70
1	9	1	3	High BK	4.15	2.52	2.95
1	9	2	3	High BK	4.15	2.52	2.97
1	9	3	3	High BK	4.15	2.52	3.56
1	9	4	3	High BK	4.15	2.52	2.85
1	10	1	4	Medium BK	2.79	2.52	2.25
1	11	2	5	Low BK	−3.56	2.52	2.02

were subsequently used as the outcome variable in a multiple-regression analysis with the person × topic BK measures used as a predictor variable. Additionally, and on the basis of the literature, two other predictor variables were also included: general language proficiency (using C-test measures), and task type.

The estimation of *R* in a regression analysis is dependent on the number of predictors (*k*) and sample size (*n*). Field (2010) recommends two rules of thumb for calculating sample size: **50 + 8k** for the overall model and **104 + k** for the predictors. There are three predictor variables in our analysis (general language proficiency, BK, and task type) bringing the recommended sample size to 74 (50 + 8x3) and 107 (104 + 3). The racked data set of 810 (81 persons × 10 topics) is well above the recommended sample size.

I carried out two separate regression analyses with the same predictor variables but with two different outcome variables: (a) person × topic speaking measures for all criteria, and (b) person × topic speaking measures for the IELTS criteria. I used a hierarchical method of data entry with the predictor variables entered in blocks. With the exception of task type, the remaining predictor variables were interval, continuous data (proficiency and BK). For task types 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:

1. Block 1: General language proficiency estimates (C-tests)
2. Block 2: BK estimates (BK questionnaires)
3. Block 3: Task type (A vs. B and A vs. C)

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, and R squared change. Regression plots, histogram of standardised residuals, and normal distribution of residuals were also specified in order to test for various other model assumptions such as homoscedasticity and heteroscedasticity. Results showed that model assumptions were met meaning that 'the model that we get for a sample can be accurately applied to the population of interest' (Field 2010:221).

The regression model summary for each step of the hierarchy as predictor variables were entered is reproduced in Table 5.33. The column labelled *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).

Table 5.33 Regression model summary (FullA)

Model	R	R^2	Adjusted R^2	SE	Change statistics					Durbin-Watson
					R^2 change	F change	*df*¹	*df*²	Sig. F change	
1	.776[a]	0.602	0.601	1.38	0.602	1218.912	1	807	0.000	
2	.794[b]	0.631	0.630	1.33	0.029	64.168	1	806	0.000	
3	.797[c]	0.636	0.634	1.32	0.005	5.059	2	804	0.007	1.233

a. Predictors: (Constant), Proficiency estimate
b. Predictors: (Constant), Proficiency estimate, BK estimates
c. Predictors: (Constant), Proficiency estimate, BK estimates, Task Type A vs Task Type B, Task Type A vs Task Type C
Dependent variable: Three-facet, five criteria (FullA)

For Model 1, $R^2 = 0.601$, which means that general language proficiency accounts for 60.1% of the variation in spoken performance on different topics. Once BK has been entered, this percentage increases to 63%, thus improving the first model by 3%. The addition of the remaining predictor variables (Model 3) increases the percentage by another 0.4. The adjusted R^2 column is an indication of the generalisability of the model; in ideal circumstances, the values for R^2 and adjusted R^2 should be very close, which is the case in this data.

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. This change in the amount of variance is explained given the F-ratio value of 1218.912, which is significant with a probability of less than 0.001. The addition of the new predictors influences R^2 to increase by 0.029 in Model 2 and 0.005 in Model 3, respectively, with associated statistically significant probability values ($p < 0.01$). The 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 (Field 2010). The value in this analysis is 1.223, which suggests that the assumption has been met.

The next output to consider is the ANOVA results (Table 5.34), which test whether 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 and highly unlikely to have happened by chance ($p < .001$). The F-ratios for the remaining models are much smaller in comparison but they are all significant at the 0.001 level, meaning that the new models (with the extra predictors) get increasingly better at predicting speaking performance on different topics.

Table 5.34 ANOVA Output (FullA)

Model		ANOVA(e)			F	Sig.
		Sum of squares	*df*	Mean square		
1	Regression	2335.702	1	2335.702	1218.912	.000[a]
	Residual	1546.388	807	1.916		
	Total	3882.09	808			
2	Regression	2449.735	2	1224.868	689.245	.000[b]
	Residual	1432.354	806	1.777		
	Total	3882.09	808			
3	Regression	2467.537	4	616.884	350.623	.000[c]
	Residual	1414.552	804	1.759		
	Total	3882.09	808			

a. Predictors: (Constant), Proficiency estimate
b. Predictors: (Constant), Proficiency estimate, BK estimates
c. Predictors: (Constant), Proficiency estimate, BK estimates, Task Type A vs Task Type B, Task Type A vs Task Type C
Dependent variable: Three-facet, five criteria (FullA)

Our main interest in this analysis is the effect of BK. As illustrated, BK accounts for the biggest improvement in predicting speaking performance (3%) after proficiency level. The ANOVA results substantiate that all predictors (including BK) significantly contribute to the model. Using the figures in the table of coefficients for Model 3 (Table 5.35), we can define our model as follows:

$$\text{Speaking performance on topic}_i = b_0 + b_1(\text{proficiency}_i) + b_2\,(\text{BK}_i) + b_3(\text{Task A vs. B}_i) + b_4(\text{task A vs. C}_i)$$
$$= -1.66 + (1.79\ \text{proficiency}_i) + (0.16\ \text{BK}_i) + (-0.03\ \text{TaskA vs. B}_i) + (-0.32\ \text{TaskA vs C}_i)$$

Table 5.35 Table of coefficients

Model		Coefficients[a]			t	Sig.
		Unstandardised coefficients		Standardised coefficients		
		B	SE	Beta		
1	(Constant)	−1.67	0.05		−33.074	0.000
	Proficiency estimate	1.837	0.053	0.776	34.913	0.000
2	(Constant)	−1.819	0.052		−34.941	0.000
	Proficiency estimate	1.787	0.051	0.755	35.005	0.000
	BK estimates	0.191	0.024	0.173	8.01	0.000
3	(Constant)	−1.664	0.08		−20.729	0.000
	Proficiency estimate	1.793	0.051	0.757	35.278	0.000
	BK estimates	0.168	0.025	0.151	6.695	0.000
	Task Type A vs Task Type B	−0.03	0.128	−0.006	−0.238	0.812
	Task Type A vs Task Type C	−0.327	0.108	−0.073	−3.027	0.003

a. Dependent variable: Three-facet, five criteria (FullA)

The *b*-values are informative in terms of the relationship between speaking performance and individual predictors. The positive values of proficiency and BK suggest that as proficiency and BK increase, speaking performance also increases. 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, $p = 0.812 > 0.05$); 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 increases by 0.16 logits. This interpretation is true when the effects of proficiency level and task type 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. In other words, BK failed to have a practical effect on speaking performance.

The above analyses were repeated for the second data set (IELTS criteria) and the model summary is presented in Table 5.36. There were no violations to the model assumptions.

Results 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. Removing the TD criterion, however, has reduced the predictive power of task type as a significant predictor of performance. BK has remained a significant predictor; however, its inclusion has only improved the predictive power of the model by 0.8% to 60.9%.

Table 5.36 Regression model summary (IELTSA)

Model	R	R^2	Adjusted R^2	SE	Change statistics					Durbin-Watson
					R^2 change	F change	*df*[1]	*df*[2]	Sig. F change	
1	.776[a]	0.602	0.601	1.393	0.602	1219.6	1	807	0.000	
2	.781[b]	0.61	0.609	1.380	0.008	16.12	1	806	0.000	
3	.781[c]	0.61	0.608	1.380	0.001	0.769	2	804	0.464	1.215

a. Predictors: (Constant), Proficiency estimate
b. Predictors: (Constant), Proficiency estimate, BK estimates
c. Predictors: (Constant), Proficiency estimate, BK estimates, Task Type A vs Task Type B, Task Type A vs Task Type C
Dependent variable: Three-facet, four criteria (IELTSA)

These results are in line with the MFRM results, demonstrating that the predictive power of BK decreases once the TD criterion is removed from the analysis. Taken together, findings from MFRM and regression analyses point to the same conclusion: topic and BK of topic have a statistically significant effect on spoken performance. This effect, however, is small and not large enough to have a practical impact on scores. BK 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.

Role of general language proficiency

Previous research and theoretical rationales advanced the possibility of an interaction between BK and general language proficiency. So far, the MFRM analyses had focused on 'main effect' models where each facet was evaluated independently in terms of its impact on measurement results. MFRM also allows for an investigation of the interaction between different facets, which is 'roughly equivalent to ANOVA' (O'Sullivan and Rignall 2007:454). In FACETS, interaction analysis is labelled *bias analysis* or differential facet analysis.

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 is then compared against observed values, which are expressed in residuals. The residuals which correspond to the specified interaction terms (e.g. BK 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, SE, size, and significance. It allows 'the effect of bias to be expressed in the same frame of reference as the element measures' (Linacre 2018a:245).

An interaction analysis – BK × language proficiency – was carried out in FACETS (Linacre 2018b) to address the following RQ: *Does BK 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 5.37 and Table 5.38, respectively. Note that both BK and language proficiency measures were divided into three groups (low, medium, high) for these analyses, as FACETS only accepts integer numbers.

Results indicate two significant bias terms; a bias z-score value larger than |2| is flagged as showing significant bias. The positive bias measure of 0.11 in Table 5.37 can be interpreted as follows: when BK levels are low, higher-proficiency examinees are at an advantage compared to those with low or medium proficiency levels. In contrast, when examinees are in the low-proficiency and low-BK group, they are at a disadvantage (bias measure = −0.1) compared to other groups. Notice also that the bias terms are not repeated for medium to high BK levels and for persons from medium proficiency levels. We can therefore deduce that having low levels of BK differentially influences persons from high and low proficiency levels. Put differently, only when BK is low, proficiency plays a role. Note also that while these bias terms are statistically significant, the size of the bias measures is small and limited to 0.1 logits. It is therefore unlikely for these bias terms to have a significant impact on performance.

Table 5.37 Summary of BK × proficiency bias analysis results (FullA)

Raw average (obs-exp)	Bias measure	Bias model SE	Bias Z-score	Bias infit MnSq	Bias outfit MnSq	BK group	Proficiency level
0.05	0.11	0.05	2.01	1.3	1.6	Low BK	High
−0.06	−0.10	0.05	−2.15	1	1	Low BK	Low

Note: obs-exp = observed minus expected.

Table 5.38 Summary of BK × proficiency bias analysis results (IELTSA)

Raw average (obs-exp)	Bias measure	Bias model SE	Bias Z-score	Bias infit MnSq	Bias outfit MnSq	BK Group	Proficiency level
0.06	0.14	0.07	2.05	1.2	1.3	LowBK	High

In IELTSA (Table 5.38), only one significant bias term is observed; the positive bias value of 0.14 suggests that when high-proficiency examinees have low levels of BK, they have a statistically significant advantage over persons with lower proficiency levels. The bias measure is small and therefore unlikely to have a substantial impact on measurement results.

These findings can be drawn on to address the RQ posed earlier; when all criteria are taken into account, there is a significant interaction between BK and language proficiency in the following cases: when BK is low, high-proficiency examinees are at an advantage compared to low-proficiency examinees. Low BK also puts persons at a low proficiency level at a disadvantage compared to examinees with higher ability levels. There was no evidence of an opposite trend, that is, high BK was not shown to favour or bias against persons from different ability levels. Despite reaching statistical significance, the bias measures were very small and therefore unlikely to have a large effect on performance. A small yet significant advantage was also observed in IELTSA for the high-proficiency examinees in relation to the low BK condition.

Topic validity evidence from a measurement perspective

In this chapter, we have focused on detailed quantitative analyses and results of the research. Let's now bring the results together to evaluate the evidence for the topic validity of the IST from a measurement perspective.

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. 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 TD 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 BK facets** were drawn upon to answer the majority of RQs. 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 and 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 to have practical significance in terms of test takers achieving 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 BK was modelled as an additional facet of the assessment context, findings suggested a significant impact of BK on performance. Differences in BK of topics were shown to pose statistically distinct levels of challenge for test takers. However, similar to the results of the topic facet, BK 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. BK was also shown to have a significant interaction with general language proficiency. In cases where BK of a topic is low, higher-proficiency test takers were found to be at an advantage compared to lower-proficiency examinees. High levels of BK on the other hand were not shown to favour or bias against persons from different ability levels.

What the findings in this chapter have shown is that both topic and BK can have a statistically significant effect on scores. Results have also illustrated how this statistical significance has systematically failed to translate into practical significance. This is because the size of these effects is shown to be negligible 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 from a measurement perspective, as scores on the speaking tasks appear to predominantly reflect the underlying speaking ability construct that the test is designed to measure. Let us now turn to the next chapter where we look at different sources of qualitative evidence for the topic validity of IST.

6 Does choice of topic matter? *A qualitative perspective*

In this chapter, we will move away from score data and consider the influence of topic and BK of topics on performance from a more qualitative perspective by analysing three additional sources of data: (a) rater interviews, (b) content of test taker speech, and (c) test taker questionnaires. While the latter consisted only of a series of closed-ended questions (analysed quantitatively with descriptive statistics), the source of data, that is, stakeholder *perceptions*, is considered qualitative and therefore presented here. Findings from the analyses of these sources serve to complement the measurement results of the previous chapter and address the following RQ: *What themes and patterns emerge from an enquiry into (a) rater perspectives, (b) the content of test taker speaking performances, and (c) test taker perspectives in relation to topic validity aspects of the speaking test under examination?*

An exploration of these different sources of data can not only generate further evidence for the topic validity argument of the IST but can also help form 'explanatory patterns' (Ellis and Barkhuizen 2005:257) for the quantitative research findings.

Insights from raters

Following the completion of the rating process and as explained in Chapter 4, I interviewed all four raters about their rating experiences. Given that they had rated approximately 200 topic-based performances each, I asked for their views on the influence of topics on speaking. All interview recordings were orthographically transcribed and thematically analysed (see Chapter 4 for more details). In the next sections, we will look at the five main emerging themes from this analysis illustrated with extracts from rater interviews.

Topic unfamiliarity: Impact on performance and ratings

All four raters in the research remarked on the adverse impact of topic unfamiliarity on features of spoken performance on the one hand and scores awarded to test takers on the other. Raters explicitly related the perceived impact to the rating scale criteria and the observation checklist. The influence of topic unfamiliarity was often associated with poor topic development (TD scale), an increase in pauses and hesitations (FC scale), a decrease in lexical diversity (LR scale) and grammatical complexity (GA scale), 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 BK whereas a negative 'knock-off effect' was predicted for ratings for the remaining scales. Illustrative rater comments are reproduced below.

> **Rater Extract 1**
> *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 2**
> *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 3**
> *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 know anything about this. It's often going to be in very simple grammar, when they know nothing about it.*
>
> **Rater Extract 4**
> *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 5**
> *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.*

Topic unfamiliarity: Impact on rater decision-making

A second theme emerging from the analysis of rater data was that topic-related issues were perceived as not only problematic for test takers but also posing a number of challenges for raters in terms of 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. The length of the response was not the only problem reported. Raters also remarked on the dilemma of rating responses that were 'short yet perfectly adequate and legitimate' or 'very long … but bearing no relation whatsoever to the question they were asked'. The TD criterion was viewed as helpful in facilitating scoring decisions for these types of responses as illustrated in the extract below.

> **Rater Extract 6**
> *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.*

Topic unfamiliarity: Impact on test taker strategies

Raters commented on distinct ways in which test takers deployed strategies in dealing with unfamiliar topics or problematic questions; for example, by going off on a tangent, speculating, waffling, or simply abandoning the topic as illustrated in the extracts below:

> **Rater Extract 7**
> *When people couldn't really talk about the question they would talk about something tangential just to be talking.*

> **Rater Extract 8**
> *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 9**
> *I mean sometimes the problems with the content ended up with people kind of waffling about something.*

Raters' attitudes towards these different kinds of strategy use were mixed, with one rater 'rooting' for those respondents who used strategies to circumvent topic-related problems whereas others questioned the use of such strategies in light of the communicative purpose of the tasks.

> **Rater Extract 10**
> *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 11**
> *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 different ways in which raters view and assess 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. Some viewed 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. Others regarded the communicative purpose of the test as equally important – if not more so – than the linguistic aspect.

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 and the strategies adopted. Lack of BK was suggested to differentially affect test takers from different proficiency levels as illustrated in the following extract:

> **Rater Extract 12**
> *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 research. The results of a bias interaction between test takers' general language proficiency and BK of topics showed that low levels of BK disadvantaged test takers from low proficiency levels compared to those with 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. Relatedly, raters also remarked on the potential negative affective influence on test takers, particularly those with lower proficiency levels catching them 'off-guard' or 'by surprise' and making them feel 'baffled' and at times scared and stressed.

Rater Extract 13
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.

Rater Extract 14
What I've noticed during the rating process is that that initial ignorance [of a topic] 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.

Test features and impact on performance

The final theme touched on by raters relates to general features of the test that might magnify the perceived impact of topic-related problems on spoken performance such as the standardised nature of the test, the examiner script, and lack of support for candidates when facing a problematic topic. These were viewed as features that could impede interaction or allow repair of problematic topic sequences. The artificial nature of the test and the lack of support is captured in the following extract.

Rater Extract 15
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 channelling 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.

One rater also drew on her teaching background to highlight the importance of supporting test takers with prompts or brainstorming activities for eliciting speech:

> **Rater Extract 16**
> *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.*

These insights align with some of the quantitative results. Task Type B topics – where prompts are presented to candidates – were found to be the least problematic from a content perspective: only three out of 121 performances flagged for qualitative analyses were related to Task Type B topics (see the next section). We can therefore argue that by providing information points for test takers, we can to some extent level the playing field in terms of BK and minimise the negative effect of topic unfamiliarity.

The findings discussed in this section have highlighted how topic unfamiliarity and lack of BK can pose challenges for test takers and raters alike with specific features of the test exacerbating the problems. Raters attributed an important role to topic unfamiliarity in negatively influencing test taker performances and scores. These findings, however, run counter to the quantitative findings of the study, as topic and BK 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 a specific test feature – the multi-question format of tasks – as a factor that can minimise the (negative) influence of lack of BK:

> **Rater Extract 17**
> *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.*

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, 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 BK on scores. What can also explain the discrepancies between the *perceived* influence of topic-related factors on scores (on the basis of rater interviews and test taker questionnaire results discussed later) 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.

Insights from content of speech

Within the field of applied linguistics, SLA, and language assessment, spoken data is predominantly used in two distinct ways: as a 'source of data' for examining learners' knowledge of an L2 and what they can do with it, and as a 'source of information' for investigating factors related to L2 learning and performance (Ellis and Barkhuizen 2005:359). In the previous chapter, participants' spoken performances were transformed into scores and used as a 'source of data' for the MFRM analyses. In this chapter, the *content* of the same spoken performances, in the form of meanings expressed by test takers, is used as a 'source of information' that can provide rich insights into topic and BK effects.

Given the large number of available spoken performances, it was not practically feasible to transcribe and analyse all data. Instead I used a sampling approach to select those performances where topic-related problems were most likely to be in effect. I drew on two additional sources – BK questionnaires and the table of unexpected responses – to inform the criteria for sample selection as follows:

Background knowledge questionnaire results. The Rasch-based BK estimates were used for extracting all performances where participants' self-reports of BK fell into the 'low BK' category. For approximately 70% of the participants, there was at least one topic for which the BK level was low. There was a total of 96 performances that were identified at this stage.

Table of unexpected responses. The data in this table flagged examples of performances where the observed score in the TD criterion was significantly lower than the expected score from the Rasch model predictions. Given that TD is the criterion most likely to absorb topic-related effects, I 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 BK effect. On this basis, a further 36 performances were identified.

There was an overlap of eight performances when the two sources were cross-checked bringing the number of selected performances to 121 (approximately 15% of the total number of performances). These speaking performances were subsequently transcribed and thematically analysed (see Chapter 4 for more details). Moreover, in line with Goetz and Le Compte's (1984) recommendation to provide instances of data that may contradict the common themes, I also looked for and will report counter-examples to some of these themes – where relevant. This type of evidence serves to 'establish the parameters or distribution of a construct' (Goetz and Le Compte 1984:175). Three main themes emerged from a thematic analysis of the content of test taker speech and we will look at each in more detail.

Topic unfamiliarity and test taker strategies

The analyses of the content of test taker performances suggested that when faced with an unfamiliar topic, test takers opt for distinct strategies; a theme that also emerged from the rater interviews. Some test takers, for example, tend to explicitly signal their lack of BK. This may be followed by complete topic abandonment where test takers fail to elaborate on a response. Questions (from tasks) and illustrative extracts (from test taker responses) are presented below:

> **Q: Where can people in your country get information about genetic research?**
>
> **Extract 18**
> TT(test taker)09: *'I don't know that where we should go and ask about anything, about genetics'*
>
> **Extract 19**
> TT017: *'Uh I don't know, I don't know if there is any genetic research in Iran, I have not heard of that'*
>
> **Extract 20**
> 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?**
>
> **Extract 21**
> 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?**
>
> **Extract 22**
> TT069: '*I have uh not enough knowledge about traditional dancing in Iran'*

Notice how some of these test takers either check comprehension by repeating the question or slightly rephrasing it in the response. This largely rules out the influence of listening (mis)comprehension in contributing to topic abandonment.

Other test takers adopt different strategies in dealing with an unfamiliar topic. Rather than disengaging from a topic or providing minimal responses, others attempt to justify their lack of BK, waffle, speculate, and/or go off on a tangent. The following extract illustrates a number of these strategies:

> **Extract 23**
> 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.'*

In this extract, the test taker repeatedly emphasises his lack of BK. By way of explanation he then refers to the unfamiliarity of topic for his 'people', that is, 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.

Problematic topics: Issues of local validity

The strongest theme emerging from the analysis of the content of test taker performances was issues related to the local validity of certain topics. Findings showed that the majority of the problematic questions or topics were related to three specific topics: Festivals, Genetic research, and Dancing. These topics required test takers to draw on previous experience or knowledge which they, as a group, did not necessarily have readily available on account of their Iranian background and other cultural and religious factors. I will explore some of these factors below with illustrative examples:

> **Q: Tell me about the most important festival in your country.**
>
> **Extract 24**
> TT050: *'Well uh as you know uh I am Iranian and in Iran actually we don't have any festivals'*
>
> **Extract 25**
> 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.'*
>
> **Extract 26**
> TT011: *'Oh, let's think of a festival. [LAUGHS]. Sorry, I really don't remember.'*

Extract 27
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.'*

Extract 28
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.

Extract 29
TT003*: 'The most (.) festival is film and I don't remember anything else, the films and somehow the football games somehow.'*

Extract 30
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.'*

Extract 31
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.'*

Q: What special food and activities are there in this festival?

Extract 32
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?

Extract 33
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 questioned the 'fairness' and local validity of this topic, in particular, in comparison to other contexts where festivals may be more popular:

> **Rater Extract 34**
> *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 35**
> *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?**
>
> **Extract 36**
> 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.'*
>
> **Extract 37**
> 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?**
>
> **Extract 38**
> 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.'*

Extract 39
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.'*

Extract 40
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.'*

The extract below suggests that the unpopularity of the topic might be tied to religious reasons:

Extract 41
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 a topic and the social, cultural, and religious backgrounds of the test takers in the study was perceptibly observable for the Festivals and Genetic research topics. The distinctive aspects of Iranian society that might render certain topics or questions irrelevant were also alluded to in relation to the topic of Dancing.

Q: Tell me about any traditional dancing in your country.

Extract 42
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.'*

Extract 43
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?]

Extract 44
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?]

Extract 45
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 how some topics may not be appropriate for particular countries and cultures; in the words of O'Sullivan (2011), they may be lacking 'local validity'. We should however be careful not to assume that these topics are problematic for *all* test takers. Some counter-examples were extracted from the data where some test takers managed to elaborate on the topics without exhibiting content-related problems:

Q: Where do people in your country find information about genetic research?

Extract 46
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 genome 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?

Extract 47
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.'*

Extract 48
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 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. It is crucial to empirically establish the degree of BK instead of relying on stereotypical notions of familiarity, which may or may not be applicable to all individuals within a group.

Topic interactions with test taker characteristics

The previous section demonstrated how some topics and/or questions might prove problematic for the majority of individuals within a sub-group of the test-taking population. This may be due to the mismatch of the topic with the social and cultural background of the test takers. The analyses also revealed how such a mismatch can also be exhibited at a more individual level where factors such as lack of personal interest in a topic or a negative affective influence of topics may inhibit performance to a certain extent. Illustrative examples are provided below:

> **Q: Tell me about any traditional dancing in your country.**
> **Extract 49**
> 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]?**
>
> **Extract 50**
> 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.'*
>
> **Q: Describe one of your friends.**
>
> **Extract 51**
> 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.'*
>
> **Q: How does water transport, like boats and ships, compare with other kinds of transport?**
>
> **Extract 52**
> 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?]**
>
> **Extract 53**
> TT017: *'Because I'm not a good dancer, and I feel a little bit stupid when I dance.'*

> **Extract 54**
> 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?]**
>
> **Extract 55**
> 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.'*
>
> **Extract 56**
> 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.'*
>
> **Extract 57**
> TT038: *'I will feel ashamed and I'm not so comfortable when I want to dance.'*

On the same topic of Dancing, the following extracts suggest a positive affective influence:

> **Extract 58**
> 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.'*
>
> **Extract 59**
> 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.'*

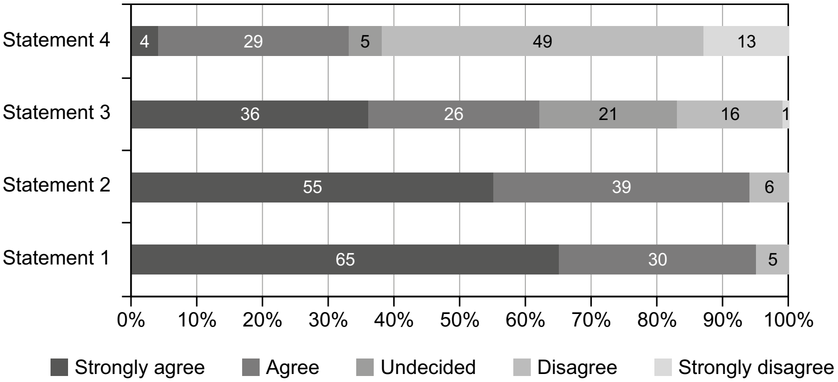
Taken together these examples illustrate the distinct ways in which test takers interact with topics and express their own individuality through the content of their speech. Relatedly, raters commented on their own engagement with the content expressed by test takers. One rater compared the speaking tests to a small window into the culture and personal lives of the test takers. Another compared the rating process to 'speed dating' where one gets to know something about different people in a very short amount of time.

Insights from test takers

The focus of this section is on test taker perspectives regarding the role of topic and BK of topic on their performance and final scores. To remind the reader, the BK questionnaire included four statements on a five-point Likert scale designed to elicit perceptions of topic effects on scores. These questions are reproduced below with the frequency of responses to each statement presented in Figure 6.1.

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.

Figure 6.1 Distribution of questionnaire responses (%); n = 82



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 for Statements 1 and 2, respectively. Only 5–6% of respondents disagreed with these statements with no respondents opting for the undecided option. Test takers' strong attitudes towards the impact of topic and BK of topic on their performance were also evident in additional comments by some of the test takers on the questionnaires: '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 the role of language proficiency is taken into account. Results display 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', the percentage has drastically dropped from 95% to 62%. Moreover, there is an increase 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).

These findings – along with rater perspectives – reveal an inconsistency between the *perceived* impact of topics and BK of topics on scores and their *observed* impact on the basis of the MFRM results discussed in Chapter 5. Test takers believe that topics and having ideas about a topic can have an effect on their final scores while the MFRM results consistently show otherwise. I will discuss these contrasting findings in more depth in the next chapter.

Topic validity from a qualitative perspective

The qualitative analyses of rater interview data and the content of test taker response data provided important insights into the impact of topic unfamiliarity on test takers and raters alike, adding more nuance for the topic validity of the test from a qualitative perspective. Findings revealed the complex ways in which characteristics of groups or individuals can interact, in distinct ways, with different topics further illustrating the inadequacy of making assumptions about test takers' levels of BK. Findings also showed how test takers and raters attribute a significant role to topic and BK of topic in influencing performance scores countering the measurement results of the score data. In the next chapter, I will bring together these different strands of findings to build a topic validity argument for the IST.

7 Building a topic validity argument

The aim of the research presented in this volume has been to examine, from a test validity perspective, the extent to which *topic* and *background knowledge* (BK) of topic have an impact on spoken performance in language proficiency interviews. The systematic variation in topics of a speaking test, exercised through the random assignment of topics to test takers, reflects two fundamental assumptions. Firstly, that topics, within a task type, are equivalent in terms of difficulty and elicit comparable performances from candidates. Secondly, that differences in the levels of BK that test takers bring to these topics as a function of group-level factors or individual test taker characteristics do not exert a systematic and significant influence on candidate performances. Evidence to the contrary would signal the presence of potential construct-irrelevant variance and test bias.

Using a mixed methods strategy of enquiry (Creswell and Plano Clark 2007), and taking the IELTS Speaking test (IST) as the research context, we have so far looked at the research problem through multiple lenses: test scores, language functions, speech content, rater behaviour and perceptions, and test taker attitudes and perceptions. My aim in this chapter is to bring together the various strands of research, synthesise the findings thematically, and to subsequently position the research in the wider literature. I will delineate both convergent and divergent findings and draw on Weir's (2005) socio-cognitive framework (SCF) of language test validation (see Chapter 1) to build a topic validity argument. The chapter will conclude with a discussion of the implications of the research and a consideration of future directions.

Topic effects on performance

The MFRM results of the study showed that at the test level, the 18 topics included in the research exhibited difficulty measures that were statistically distinct. This was as expected; the IST is designed to include task types of differing difficulty levels. Focusing at the task level, results suggested that topics within each task type of the IST, i.e. Part 1 (Task Type A), Part 2 (Task Type B) and Part 3 (Task Type C) could be divided into a minimum of two statistically distinct difficulty levels. In other words, topics within each task type could not be considered equivalent or 'parallel'. 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 rating criteria. Findings showed that these differences were not large enough to translate into differences in test takers' speaking scores, as the minimum speaking ability required to move across band scores consistently exceeded the maximum difference between topic difficulties, both at task and test levels. These results demonstrate the absence of a practical effect of differences in topic difficulties on performance scores; an encouraging finding for the test developers, serving as one element of topic validity evidence that justifies, on the basis of the score data, the random assignment of topics to test takers in the context of the IST.

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 adapted version of the observation checklist (O'Sullivan et al 2002). Illustrative comparisons of observed functions across two topics within each task type revealed both similarities and differences in the range of functions elicited by topics. These variations imply that topics, which are seemingly equal, may tap into different aspects of the underlying speaking construct and as such, cannot be considered parallel. Put differently, 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 et al 2002). Two important caveats, however, need to be borne in mind. Firstly, in the operational IST, a given score represents performance on the three test parts comprising five topics. 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'. An in-depth analysis of these issues was beyond the scope of this research and therefore the findings are not conclusive. Further research however is needed to investigate topic effects on qualitative aspects of test takers' spoken performances.

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. Fulcher (2003) and Fulcher and Márquez Reiter (2003) 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; emphasis in original). The current findings closely align with Fulcher's (2003) argument; while the results of the analyses of functions suggest qualitative differences in the language produced by test takers are attributable, in part, to the task condition of topic, the MFRM analyses of score data illustrate how differences in topic difficulty measures have failed to have a practical effect on test scores.

Background knowledge effects on performance

The measurement results of the study generally suggested a statistically significant effect for BK with a small effect size, failing to reach practical significance. An emerging theme from the analyses of rater transcripts was the *perceived* strong impact of topic unfamiliarity on the language produced by test takers. We can therefore observe both parallels and contrasts from these two data sources; the results converged in respect of BK exerting an impact on performance but diverged in relation to the *degree* of this impact. The MFRM results for the three BK conditions (low, medium, and high) showed a statistically significant main effect of BK where low levels of BK – suggesting topic unfamiliarity – were shown to consistently and systematically pose the greatest level of challenge for test takers. This is compatible with rater observations. High levels of BK were also shown to have a facilitative effect on performance. These are in line with several studies in the literature that have found a statistically significant role for BK in L2 performance (e.g. He and Shi 2012, Huang et al 2018, Krekeler 2006, Schmidt-Rinehart 1994, Tedick 1990). The findings, however, run counter to those reported in Jennings et al (1999) and Lee and Anderson (2007) where BK, 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 BK 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 this research can therefore be a reflection of the independent nature of the speaking tasks (particularly Task Types A and C), which require test takers to largely rely on their own BK in responding to questions.

The main point of divergence in the findings is the extent to which statistical significance translates into practical significance, that is, in terms of influence on achieving higher or lower band levels. The analyses of test taker questionnaire data indicate that the majority of test takers (95%) place a great importance on both topic and BK variables as factors that can affect their final scores. This finding echoes the results of other empirical research examining test taker perceptions of topic and BK effects on scores (e.g. He and Shi 2012, Jennings et al 1999). The raters in the study also attributed an important role to BK, or lack thereof, in shaping features of test takers' performances on the one hand, and scores awarded to test takers on the other. In relating performances to the rating scale criteria, the influence of topic unfamiliarity was often perceived to be 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 noted. Pronunciation (P scale) was considered the scale least likely to be negatively affected by lack of BK. In contrast, an adverse effect on ratings for the remaining scales was predicted.

The importance attributed to topic and BK in influencing scores from the perspectives of test takers and raters was not reflected in the measurement results of the study. The statistically significant difference between different BK measures failed to translate into a 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. It therefore appears that perceptions regarding the magnitude of the BK effect may not be necessarily reflected in score data.

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 – limited to 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 fulfilment and relevance criterion' for those topics which males were assumed to have higher BK of (Lumley and O'Sullivan 2005:432–433). These findings 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 whereas general language proficiency exhibited 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 rater perceptions in the current study). The small and negligible effect sizes, however, 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). The general trend that BK has a statistically significant but practically negligible influence on performance scores is supported in the present data.

We can therefore observe an apparent mismatch between the perceived (strong) influence of topic and BK on speaking scores – as voiced by raters and test takers – and the absence of a practical influence of these variables on scores based on the measurement results. Let us now consider possible explanations for this disparity. One possible explanation relates to specific features and constraints of the IST and the absence of a built-in support mechanism for dealing with topic-related problems. This was one of the emerging themes from the qualitative analyses of examiner transcripts. Another theme touched on the format of the speaking tasks as potentially mediating the effects of BK of topic. Building on these themes, it can be argued that certain features of the test and test tasks can magnify the salience of topic-related effects while simultaneously moderating the (negative) impact of lack of BK. Let me elaborate.

Two important features of the IST in relation to topics were outlined early on in Chapter 1: 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, we can postulate that the combination of these two features amplifies the perceived impact of topic/BK of topic in two 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 BK 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 does arise in the test, neither test takers nor examiners feel that there is room for addressing the problem (Seedhouse 2018, 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 of the interview. The standardised nature of the test may therefore leave both parties with little to fall back on when the intended purpose of the test – generating samples of speech through topics – is thwarted by the influence of topic unfamiliarity.

A second possible explanation alluded to by the raters in the study pertains to the potential (negative) emotional impact of topic unfamiliarity on test takers giving rise to feelings of anxiety, confusion, or apprehension. These can leave a strong affective impression on test takers (Bachman and Palmer 1996) and thus explain the perceived prominence of topics and BK of topics in the eyes of test takers.

I now propose three reasons for the absence of a strong BK effect despite its perceived salience in influencing scores as follows:

(a) the inclusion of the TD criterion. The intention behind the use of the TD scale in this research was to capture non-linguistic and content-oriented features of test takers' performances in respect of the development of ideas. The effects of BK were therefore likely to have been absorbed by this criterion thus reducing or minimising influence on the remaining criteria. Evidence in support of this argument derives from the measurement results where TD was shown as the only scale likely to be strongly affected by differences in BK levels.

(b) the multi-question, multi-task format of the speaking tests. Insights from rater remarks and the analyses of the content of test taker performances illustrated how it is not necessarily a topic that might be unfamiliar or problematic but rather, specific questions within a topic sequence that might require test takers to draw on BK that is not readily available to them. This observation is in line with findings by Seedhouse and Harris (2011) in relation to problematic questions in the IELTS speaking tasks. BK might therefore 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 therefore fades within the broader performance.

(c) test-taking strategies. As evidenced in rater remarks and illustrative examples from the content of performances, some test takers circumvent

BK-related obstacles by deploying a number of different test-taking strategies such as speculating, waffling, or going off on a tangent, which allow them to generate speech regardless of topic unfamiliarity. Raters' perspectives on the use of such strategies were mixed; on the one hand, the speech-generating function of these strategies was viewed positively in facilitating the rating of linguistic features of performance. The use of strategies such as waffling or going off on a tangent, on the other hand, was perceived negatively due to test takers' failure to address the task adequately. One rater raised a fundamental issue related to the purpose of communication and whether there is any real 'coherence' in a response that does not answer a question, no matter how extended that response is. These findings suggest that topic unfamiliarity may instigate different test-taking strategies that can potentially minimise the impact of lack of BK on linguistic aspects of performance though not necessarily on the fulfilment of the task in terms of topic development.

Role of general language proficiency

The different strands of research generally converge in relation to the role of general language proficiency, its interaction with BK, and their respective contributions to overall performance. Findings from rater interviews, for example, suggested that lack of BK might differentially affect test takers from high and low proficiency levels. This observation was independently and empirically addressed in the MFRM analyses. The results of a bias analysis between BK and proficiency indicated that when BK 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 did not advantage/disadvantage examinees from different proficiency levels. Insights from rater remarks shed light on these findings: low-proficiency examinees are more likely to be negatively influenced, on an emotional level, by unfamiliar topics, leading them to abandon the topic or disengage from the question whereas higher-proficiency test takers may be able to draw on a variety of strategies to deal with problematic topics.

The test taker questionnaire responses partially aligned with the above findings; this was evidenced in the shift in their pattern of responses when they were asked to express their views on influence of topic and BK of topic on scores. The majority of participants (95%) agreed that choice of topic and having ideas about a topic might affect their final scores. However, when the same statement included a proficiency element, that is, 'choice of topic is not important if my English is good enough' about one-third of the participants (33%) agreed with the statement. We can therefore infer that from the perspective of some test takers, topic and BK are less likely to exert an influence at higher levels of proficiency.

Earlier I touched on the relative contribution of BK and general language proficiency in explaining variance in spoken performance. 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 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. In light of the strong positive correlations between C-test measures and speaking ability measures, we can argue that higher-order constructs such as general language proficiency and/or speaking ability are the main determinants of performance in the IST. These findings run counter to the conclusions drawn from He and Shi (2012:460) in respect of 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'. Results, however, 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'.

In sum, general language proficiency and oral speaking ability were shown to be the main determinants of performance in this research and the practical influence of the construct-irrelevant variable of BK of topic on speaking performance scores was found to be minimal. These serve as important pieces of topic validity evidence for the speaking test under examination.

Interaction of background knowledge and topic difficulty

The fourth set of findings from the research relates to the interaction between topic difficulty and BK levels where the quantitative and qualitative strands of enquiry intersected once again. An examination of the influence of BK on the measurement results of other facets revealed that, with the exception of the topic facet, BK had effectively no impact on other facets. The BK influence on topic measurement results was manifested in three ways: firstly, in a shift in the rank ordering of different topic elements; secondly, in changes to topic difficulty measures; and thirdly, in variations in the overall range of topic difficulty measures. Distinct from the other facets, the measurement results of the topic facet fluctuated with the inclusion of the test taker characteristic of BK of topic. These fluctuations were not substantial; nevertheless, the finding that the measurement results of the topic facet did not remain stable is noteworthy in lending strong support to Bachman's (2002:464) argument discussed in Chapter 2 regarding the problematic notion of conceptualising '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 level of BK test takers bring to the topic. This was particularly striking for two of the topics – Festivals and Genetic research – which were estimated as the most difficult topics within their respective task types (A and C) when BK was not explicitly parameterised. When BK was modelled in the MFRM analyses, however, 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 BK in this particular sample of test takers that interacted with the topic, resulting in higher difficulty indices.

The thematic analysis of the content of test taker performances substantiated and shed further light on the measurement 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 own context suggesting lack of local validity of certain topics (O'Sullivan 2011).

A closer examination of several question–answer sequences displayed how unfamiliarity (of topic or a specific question) resulted in candidates closing a question down by providing short responses such as 'I don't know' or 'I'm not really interested in that'. One of the raters in the study reported facing a dilemma in rating such performances where responses are 'perfectly adequate' in light of the test takers' lack of BK but unacceptable nonetheless in terms of topic development. This finding is in line with Seedhouse and Harris' (2011) observations in respect of the primacy of the question–answer component to the topic component in speaking tasks on those occasions where the two components do not coincide, leading the authors to contend that 'candidates can answer questions without developing topics' (Seedhouse and Harris 2011:73). I would therefore like to argue that lack of BK or topic unfamiliarity may increase the likelihood of candidates responding to questions without necessarily developing or elaborating on topics, which arguably counters their intended speech-generating function.

Another parallel can be drawn between the qualitative findings of the current study and those by Seedhouse and Harris (2011:105) in relation to problematic questions or topics within the context of the IST:

> 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.

Results from the analysis of the content of test taker performances illustrated how a seemingly familiar topic such as 'Festivals' became a source of 'unforeseen trouble' (Seedhouse and Harris 2011:105) for most of the Iranian test takers in this study. Contrasting extracts, on the other hand, demonstrated how within the same local context, some test takers managed to elaborate on the problematic topics without exhibiting content-related problems. These findings reveal how a given topic can potentially bias against a particular group of test takers but simultaneously 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 thematic analyses also exemplified the intricate ways in which test takers engage with a topic drawing on their previous experiences and personal, cultural, and religious backgrounds, highlighting 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'.

Role of the topic development criterion

To remind the reader, the topic development (TD) criterion was included in this study as means of isolating the content-related effects of topic and BK of topics on performance scores. All MFRM analyses were therefore carried out both with and without the TD criterion. Here, I will bring together the different sets of findings from these analyses to evaluate the contribution of the criterion within the assessment context.

The MFRM results showed that TD was the easiest of all five criteria, exhibiting a markedly lower difficulty level, meaning that test takers have 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 BK of topics implied that the scale was functioning as intended in absorbing the non-linguistic and content-oriented features of performances. 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 comparatively high, were within an acceptable range. These indices indicated that the criterion was contributing, as an independent criterion, to the underlying speaking construct and explaining some of the variance in the speaking scores. An examination of the table of unexpected responses showed that where the criterion was associated with large residuals, it was mostly related to low-ability examinees achieving scores higher than expected on the TD criterion and conversely, high-ability examinees achieving scores lower than expected. Taken together, the results 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.

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, BK of topics) on scores were shown to be negligible 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.

Towards building a topic validity argument

In this section, I will draw on Weir's (2005) SCF of language test validation to bring together the various strands of findings in relation to each of the main elements of the framework and build a cohesive topic validity argument for the IST. To remind the reader, SCF consists of six central elements: test taker characteristics, cognitive validity, context validity, scoring validity, consequential validity, and criterion-related validity (see Chapter 1 for a more detailed account of the SCF and its elements). My aim here is 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 speaking 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. These measurable differences, however, failed to reach 'practical significance' (Fulcher 2003). In the words of Dorans and Feigenbaum (1994), these are differences that 'do not matter', 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. We can therefore argue 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). Topics within each task type were also compared in respect of 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 has 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 statistically significant ways in which test takers' level of BK of topics exerted an influence on performance measures. The qualitative analyses of the content of spoken performances also revealed a complex interaction between topics and various test taker characteristics such as 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 BK levels, despite statistical significance, failed to have a practical impact on scores: the maximum level of difference between different BK levels consistently fell below the minimum speaking ability required to move across adjacent band scores and across different criteria.

In terms of **cognitive validity**, differences in topic difficulty measures can provide indirect evidence for varying the levels of cognitive demand put on the test takers as a function of topic familiarity and abstractedness of tasks. The sequencing of topics, however, did not always follow the intended progression in cognitive demand from easy to difficult, or from familiar to abstract. As the study's results have illustrated, the nature of BK is highly individual and test taker dependent, which makes it difficult, if not impossible, to make *a priori* assumptions of familiarity. This might lead to inadequate categorisations 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 exhibited high levels of consistency within themselves although exercised significantly different severity levels compared to each other. By adopting an MFRM approach, however, rater severity was directly parameterised 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 BK of topic on scores revealed that, despite significant differences in the element measures of each facet, differences were too small to have a practical effect on scores. Based on this evidence, we can argue 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 – used as a measure of general language proficiency – lend further evidence of criterion-related validity.

The last element[1] to consider is the **test taker**. 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 given their cultural background. This suggests an element of topic-related bias, which can bring 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, and personal interest (amongst others) were also in interaction with the topics.

[1] Note that the scope of the study did not allow for the collection of consequential validity evidence.

MFRM results demonstrated empirically that different levels of BK 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 BK 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 BK was associated with negative affective influence on test takers, particularly lower-proficiency individuals. Taken together, these findings raise a number of fairness and validity concerns in relation to test topics which I will return to shortly.

Broadly speaking, the results of the study lend strong support to the topic validity of the speaking test under examination; the quantitative findings of the study have consistently shown that large differences in topic measures and test takers' BK of topics fail to have a practical impact on test scores. Both quantitative and qualitative strands of the study indicate that the random assignment of topics and test takers' level of BK 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 this study have rejected the absence of bias but have nevertheless shown that any bias present is not large enough to dramatically 'distort' test results 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 of the influence of topic and BK on test takers, raters, features of performance, and raters' decision-making should be ignored. As famously put by Messick (1989:13), 'validity is a matter of degree, not all or none'. The validity issues and concerns raised in this research need to be considered and steps should be taken to increase the topic validity of tests, which brings me to the implications of the research.

Implications

The main implication from this research is that in the speaking performance assessment context of the IST (and tests that are similar in format and design), the topic of the performance tasks and the level of BK that test takers bring to the topics are *unlikely* to have a large practical 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 – was shown not to be large enough to distort test results. Attempts should nevertheless be made to reduce (any) negative impact of topics and BK of topics. Findings also hold further methodological, theoretical, and practical implications for speaking test design and research on oral performance assessment.

Methodological implications

The research has two important methodological implications. Firstly, the study has contributed to the body of research on the various advantages of using the Rasch family of models in examining the quality of various measurement instruments employed for the research (Tennant and Conaghan 2007). In using MFRM (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. These findings attest to the usability of the Rasch family of models in L2 performance assessment contexts and in instrument design and validation. Secondly, the research has shed light on the benefits of adopting a mixed methods strategy of enquiry in LT research (Moeller, Creswell and Saville (Eds) 2016) in deepening our understanding of the phenomenon under examination. In bringing the results from score data, the functions checklist, and 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 or divergences in findings.

Theoretical implications

Findings from the research have several theoretical implications in relation to facets of speaking performance assessment, test performance models, conceptualisations of task difficulty, and speaking test construct definition.

The study has contributed to the body of scholarly work exploring the impact of various facets of the assessment context on performance, focusing specifically on two facets of topic and BK of topic. Results suggested a small yet statistically significant impact of both variables on performance while also revealing complex interactions between various elements of the assessment context, lending 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. Eckes 2009, McNamara 1996, O'Sullivan and Weir 2011, Weir 2005).

Another important finding pertained to the conceptualisation of the notion of task or topic 'difficulty'. Results of the research 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, I would like to argue that the notion of 'parallel or equivalent' tasks or topics may be misleading. This argument can be extended to explain why research endeavours in predicting task difficulty (Norris et al 2002, Skehan 1998) on the basis of task characteristics alone may have been inconsistent or failed to reach conclusive findings. In line with Bachman (2002), it is important to make a distinction between characteristics of the task and those of the test takers and to conceptualise difficulty as interactions between the two.

The final theoretical implication of the research pertains to the definition of the speaking construct underlying the test. As discussed throughout the volume, a TD or task fulfilment criterion is not currently part of the IELTS rating scale but for the purposes of this study, a TD criterion was included as a 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'.

Findings from my study align with Sato (2012) and Elder et al (2017) who criticise current models of communicative competence for over-reliance on linguistic features and call for the inclusion of more complex non-linguistic features. The addition of a content-oriented criterion can serve to expand the speaking test construct in IELTS with an increased emphasis on the successful communication of meaning rather than focusing solely on the linguistic quality of performance (Elder et al 2017, McNamara 1996, Sato 2012). This can also lead to positive washback in the classroom by aligning assessment criteria with communicative language teaching practices. A disadvantage, however, might be the increase in cognitive demand on raters, as they would have to assign scores on an additional criterion and for each individual task. Additional support and training would be needed for raters to not only score content-related aspects of speech but to also deal with off-topic or rehearsed responses. Moreover, the inclusion of such a criterion would run the risk of magnifying the impact of BK on performance and therefore necessitate mechanisms for minimising the (negative) impact of BK and/or providing separate weightings for the different criteria.

There is clearly a balancing act between theoretical stances in defining constructs and the practical demands of large-scale assessment: the absence of a content-oriented criterion risks construct-underrepresentation whereas its inclusion can have myriad practical implications. Any decisions would therefore need to be made in light of both sets of considerations.

Practical implications and recommendations

Below are some of the practical implications of the research for speaking performance assessment along with suggestions and recommendations.

Minimising (negative) BK effects

Results of the research related to the interaction of topic of the tasks and the test takers' BK of topics suggest that it is very difficult, if not impossible, to establish topic difficulty and/or familiarity *a priori*. As Kasper and Ross (2007:2,065) 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, for the most part, a process of expert judgement followed by 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 decidedly individual way in which test takers interact with topics circumvents the possibility of making generalisations regarding topic familiarity/difficulty. The first practical implication of the study is a suggestion to shift some of these efforts and focus instead on minimising any negative impact of BK, ensuring that mechanisms are in place for dealing with problematic topics and/or the lack of BK while maximising the opportunities to speak. Some possibilities include:

(a) 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 BK is reduced. Secondly, by giving test takers a choice, they are given autonomy (Kenyon and Malone 2010) and agency, 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 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. These advantages would once again need to be weighed against practical considerations related to allowing an additional time window for topic review and selection by test takers that would increase overall test length.

(b) Levelling the playing field. Another possible approach for reducing the negative impact of BK is to provide test takers with the necessary information to respond to tasks, for example in the form of information-based prompts (similar to Task Type B). Another approach is to move away from topic-driven independent speaking tasks and include a broader range of speaking tasks such as integrated tasks[2] that do not necessarily require test takers to draw solely on their own BK.

(c) Flexibility in the use of the examiner script. The analyses of rater interview transcripts and recent IELTS research (Inoue et al 2021, Seedhouse and Nakatsuhara 2018) have highlighted some of the constraints posed by the examiner script in not allowing them the flexibility to deal with problematic interactions and occasions where topic familiarity may hinder performance. Extending Linacre's (2018a) argument for viewing raters as 'independent experts' and not 'scoring machines', a more flexible script or interlocutor frame may equally allow examiners to act as experts rather than test delivery machines, helping them deal with problematic topic sequences more effectively.

Consideration of different marking models

The second practical implication of the research pertains to the TD criterion. Should a TD criterion be included in a speaking test within a multi-task test format, then scores – at least on this specific criterion – 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 the cognitive demand posed on examiners and/or raters (for a recent discussion on different marking models see Khabbazbashi and Galaczi 2020). Considerations should therefore be

[2] 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).

given to alternative marking models that can both support the inclusion of the criterion but also be feasible within the practical restraints of the assessment context.

Advice for test takers

The absence of a practical effect of topic and BK 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 effect 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 noted by Seedhouse and Harris (2011), answering questions does 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.

A note on limitations

This volume has endeavoured to contribute to a better understanding of the role of topic and BK of topics in L2 performance assessment contexts. There are, however, several limitations that need to be taken into account with regard to the scope and design of the study.

A general limitation of the research is its specific context, limited to a particular test of speaking with particular features, which can reduce the generalisability of findings to other test settings. Furthermore, the research data was not from live exam conditions and the experimental settings of the study diverged to some extent from operational exam settings; for example, by employing non-IELTS raters, marking the test by task, and including a TD criterion, which can limit the direct application of the study's findings to the specific test under examination. Other more specific limitations are detailed below:

1. 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 BK on performance scores. This 'indifference threshold' (Dorans and Liu 2009:13) is defined in terms of the score units of a 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.

2. 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.
3. 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 magnify the impact of topic and (lack of) BK of topic on performance.
4. 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 BK in order to better examine the sources of variations in observed functions.
5. The MFRM analyses involved the mathematical – and not experimental – inclusion and removal of the TD criterion in the analyses in order to examine the extent to which topic and BK of topic affected performance scores on the remaining criteria (FC, LR, GA, P) once TD was eliminated. It is therefore not possible to predict the influence of topic and BK on scores had the TD not been included in the scales in the first place.

Notwithstanding these limitations, I hope that my research has contributed to a better understanding of the influence of topic and BK in performance assessment contexts and helped provide empirical evidence for addressing important validity concerns in speaking tests.

Concluding remarks

As I write this final chapter in the spring of 2021, it has been 20 years since the last major revision to the IST and the world is going through a global pandemic that has fundamentally changed the ways in which we communicate. There has been a paradigm shift towards multimodal communication (Herring 2018), more emphasis is placed on non-linguistic and content-oriented aspects of communication (Elder et al 2017), and the affordances of video-conferencing technologies have increased access and provided more opportunities for online interactions.

The time is therefore ripe for test developers to build on previous research and utilise new technologies to create a new generation of assessments that reflect our changing world and are fit for purpose. The research covered in this volume generally lent strong support to the topic validity of the IST but it also highlighted several areas that could be further improved such as an over-reliance on independent speaking tasks and the absence of a content-oriented criterion. The next generation of IELTS might therefore see a move away from independent tasks towards the integration of skills and allowing more agency to test takers, greatly facilitated in online settings. There might be a stronger focus on content-oriented aspects of communication with BK not as something to be ignored or controlled for but rather re-conceptualised as part of the language ability construct (Banerjee 2019, Purpura 2016). With the wide use of automated scoring technologies, we might see hybrid approaches to marking (De Jong 2018, Isaacs 2018) with some features of speech such as fluency and pronunciation marked by machines while higher-level aspects of language use such as topic development or task achievement can be scored by human raters, capitalising on their respective strengths.

Possibilities are endless, with each bringing an array of new questions, challenges, and research avenues. Exciting 'topics' await and our role as researchers and test designers is to ensure that all emerging possibilities are continuously evaluated in light of 'the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other models of assessment' (Messick 1989:13; emphases in original).

Appendix A 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
	Festivals	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
	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, lake or sea	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
	Describe an important choice	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
	The economic importance of rivers, lakes and the sea	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
	Important choices	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
	Choices in everyday life	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 et al (2006:129) and Weir and Wu (2006).

Appendix B Speaking task topics

Task Type A topics

Topic A.1: Family (UCLES 2002:29) **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?	**Topic A.2: Leisure time** (UCLES 2005:80) **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 in your city? • How do you usually spend your holidays? • Is there anywhere you would particularly like to visit? [Why?]
Topic A.3: Festivals (UCLES 2002:53) **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?	**Topic A.4: Colour** (UCLES 2006:54) **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?

Topic A.5: Keeping in contact (UCLES 2009:32) **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?]	**Topic A.6: Dancing** (UCLES 2008:32) **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?]

Task Type B topics

Topic B.1: Describe a friend (UCLES 2006:77) 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.	**Topic B.2: Describe a river, lake or sea** (UCLES 2005:80) 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.
Topic B.3: Describe someone in your family (UCLES 2008:32) 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.	**Topic B.4: Describe an important choice** (UCLES 2008:100) 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.

Task Type C topics

Topic C.1: Qualities of friends (UCLES 2006:77) • 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** (UCLES 2006:77) • 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?]
Topic C.3: Water-based leisure activities (UCLES 2005:80) • 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** (UCLES 2005:80) • 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?]
Topic C.5: Family similarities (UCLES 2008:32) • 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.6: Genetic research** (UCLES 2008:32) • Where can people in your country get information about genetic research? • How do people in your country feel about genetic research? • Should this research be funded by governments or private companies? Why?

Topic C.7: Important choices	**Topic C.8: Choices in everyday life**
(UCLES 2008:100)	(UCLES 2008:100)
• 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?	• 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 than in the past?

Appendix C Speaking test forms

Speaking Test Form W

PART I (Task Type A)

Topic A.1: Family (UCLES 2002:29)

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?

Topic A.2: Leisure time (UCLES 2005:80)

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 in your city?
- How do you usually spend your holidays?
- Is there anywhere you would particularly like to visit? [Why?]

PART II (Task Type B)

Topic B.1: Describe a friend (UCLES 2006:77)

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.

PART III (Task Type C)

Topic C.1: Qualities of friends (UCLES 2006:77)

- 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 (UCLES 2006:77)

- 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?]

Speaking Test Form X

PART I (Task Type A)

Topic A.3: Festivals (UCLES 2002:53)

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?

Topic A.4: Colour (UCLES 2006:54)

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?

PART II (Task Type B)

Topic B.2: Describe a river, lake or sea (UCLES 2005:80)

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.

PART III (Task Type C)

Topic C.3: Water-based leisure activities (UCLES 2005:80)

- 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 (UCLES 2005:80)

- 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?]

Speaking Test Form Y

PART I (Task Type A)

As Part I Speaking Test Form W.

PART II (Task Type B)

Topic B.3: Describe someone in your family (UCLES 2008:32)

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.

PART III (Task Type C)

Topic C.5: Family similarities (UCLES 2008:32)

- 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.6: Genetic research (UCLES 2008:32)

- Where can people in your country get information about genetic research?
- How do people in your country feel about genetic research?
- Should this research be funded by governments or private companies? Why?

Speaking Test Form Z

PART I (Task Type A)

Topic A.5: Keeping in contact (UCLES 2009:32)

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?]

Topic A.6: Dancing (UCLES 2008:32)

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?]

PART II (Task Type B)

Task B.4: Describe an important choice (UCLES 2008:100)

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.

PART III (Task Type C)

Topic C.7: Important choices (UCLES 2008:100)

- 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.8: Choices in everyday life (UCLES 2008:100)

- 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 than in the past?

Appendix D C-tests

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 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, www.english-online.at/society/nobel-prize/nobel-prize.htm)

Text: Lack of sleep (Text 2)

For many people, lack of sleep is rarely a matter of choice. Some ha_____[1] problems get_____[2] to sleep, oth_____[3] with sta_____[4] asleep un_____[5] the mor_____[6]. Despite pop_____[7] belief th_____[8] sleep i_____[9] one lo_____[10] event, rese_____[11] shows th_____[12], in a_____[13] average ni_____[14], there a_____[15] five sta_____[16] of sl_____[17] and four cyc_____[18], during wh_____[19] the sequ_____[20] of sta_____[21] is repe_____[22].

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 fi_____[1] the scien_____[2] language o_____[3] choice rema_____[4] Latin. It all_____[5] scientists t_____[6] communicate wi_____[7] other soci_____[8] privileged thin_____[9] while prote_____[10] their wo_____[11] from unwa_____[12] exploitation. Some_____[13], the de_____[14] to protect id_____[15] se_____[16] to ha_____[17] been stro_____[18] than t_____[19] desire t_____[20] communicate th_____[21], particularly i_____[22] the ca_____[23] of mathematicians and doc_____[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: 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 evi_____^{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)

As C-test Version I Text 1 above.

Text: Street art (Text 6)

Street art is a very popular form of art that is spreading quickly all over the world. You c_____^{1} find i_____^{2} on buil_____^{3}, sidewalks, str_____^{4} signs a_____^{5} trash ca_____^{6} from Tokyo t_____^{7} Paris. Street art has bec_____^{8} a global cul_____^{9} and ev_____^{10} art muse_____^{11} and gall_____^{12} are colle_____^{13} the wo_____^{14} of street art_____^{15}. Street art started o_____^{16} very secr_____^{17} because i_____^{18} is ill_____^{19} to pa_____^{20} on public a_____^{21} private prop_____^{22} without permi_____^{23}. People of_____^{24} have diff_____^{25} opinions ab_____^{26} street art. So_____^{27} think i_____^{28} is a cr_____^{29} and oth_____^{30} think i_____^{31} is a ve_____^{32} beautiful, n_____^{33} form of culture.

Score: _____/33 items
(Adapted from English-Online, n.d.)

Text: Minority languages (Text 7)

There are currently approximately 6,800 languages in the world. This gr_____[1] variety o_____[2] languages ca_____[3] about lar_____[4] as a res_____[5] of geogra_____[6] isolation. B_____[7] in tod_____[8] world, fac_____[9] such a_____[10] government initi_____[11] and econ_____[12] globalization a_____[13] contributing t_____[14] a huge decr_____[15] in t_____[16] number of languages. O_____[17] factor wh_____[18] may he_____[19] to ens_____[20] that so_____[21] endan_____[22] languages d_____[23] not d_____[24] out compl_____[25] is peo_____[26] increasing apprec_____[27] of th_____[28] cultural iden_____[29]. This has been encouraged through programs of language classes for children and through 'apprentice' schemes.

Score: _______/29 items
(Adapted from UCLES 2004:44)

Text: Taking a gap year (Text 5)

As C-test Version I Text 5 above.

Text: Children's books (Text 8)

Everyone has a favourite children's book. Fond memo_____[1] remain o_____[2] books th_____[3] we ha_____[4] read a_____[5] re-read, transp_____[6] the rea_____[7] back t_____[8] childhood. N_____[9] matter i_____[10] they've lo_____[11] the ha_____[12] of rea_____[13] in la_____[14] life; i_____[15] is a ra_____[16] adult w_____[17] does n_____[18] have a resi_____[19] tenderness f_____[20] a defining wo_____[21], whether it's *The Story of Babar, A Bear Called Paddington* or *Ballet Shoes*. Some_____[22] it's the te_____[23] that attr_____[24] but mo_____[25] often th_____[26] not, it's the illust_____[27] that dr_____[28] in the young reader.

Score: _______/28 items
(Adapted from *Time Out* 2007)

C-test Version III

Text: The Nobel Prize (Text 1)

As C-test Version I Text 1 above.

Text: Street art (Text 6)

As C-test Version II Text 6 above.

Text: Noise (Text 9)

In general, it is plausible to suppose that we should prefer peace and quiet to noise. And y_____^{1} most o_____^{2} us ha_____^{3} had t_____^{4} experience o_____^{5} having t_____^{6} adjust t_____^{7} sleeping i_____^{8} the moun_____^{9} or t_____^{10} countryside bec_____^{11} it w_____^{12} initially 't_____^{13} quiet': a_____^{14} experience th_____^{15} suggests th_____^{16} humans a_____^{17} capable o_____^{18} adapting t_____^{19} a wide ra_____^{20} of no_____^{21} levels. Research supp_____^{22} this vi_____^{23}.

Score: _______/23 items
(Adapted from UCLES 2009:96)

Text: Language in science (Text 3)

As C-test Version I Text 3 above.

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)

References

Alderson, J C and Urquhart, A H (1983) The effect of student background discipline on comprehension: a pilot study, in Hughes, A and Porter, D (Eds) *Current Developments in Language Testing*, London: Academic Press, 121–127.

Alderson, J C and Urquhart, A H (1985) The effect of students' academic discipline on their performance on ESP reading tests, *Language Testing* 2, 192–204.

Anastasi, A (1988) *Psychological Testing* (Sixth edition), New York: Macmillan.

Andrich, D (1978) Application of a psychometric rating model to ordered categories which are scored with successive integers, *Applied Psychological Measurement* 2, 581–594.

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, 95–104.

Andrich, D (1985) A latent-trait model for items with response dependencies: Implications for test construction and analysis, in Embretson, S (Ed) *Test Design: Contributions from Psychology, Education and Psychometrics*, New York: Academic Press, 245–273.

Andrich, D (2004) Controversy and the Rasch model: a characteristic of incompatible paradigms?, *Medical Care* I7–I16.

Andrich, D and Hagquist, C (2012) Real and artificial differential item functioning, *Journal of Educational and Behavioral Statistics* 37, 387–416.

Andrich, D and Marais, I (2010) *Introduction to Rasch Measurement of Modern Test Theory*, Perth: University of Western Australia.

Andrich, D, De Jong, J and Sheridan, B E (1997) Diagnostic opportunities with the Rasch model for ordered response categories, in Rost, J and Langeheine, R (Eds) *Applications of Latent Trait and Latent Class Models in the Social Sciences*, Münster: Waxmann Münster, 59–71.

Andrich, D, Lyne, A, Sheridan, B E and Luo, G (2010) *RUMM 2030*, Perth: RUMM Laboratory.

Babaii, E and Ansary, H (2001) The C-test: a valid operationalization of reduced redundancy principle?, *System* 29, 209–219.

Bachman, L F (2002) Some reflections on task-based language performance assessment, *Language Testing* 19, 453–476.

Bachman, L F (2004) *Statistical Analyses for Language Assessment*, Cambridge: Cambridge University Press.

Bachman, L F (2011) *How do different language frameworks impact language assessment practice?*, presented at the ALTE 4th International Conference, Kraków, Poland, July 2011.

Bachman, L F and Palmer, A S (1996) *Language Testing in Practice: Designing and Developing Useful Language Tests*, 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, 238–257.

Baker, B A (2012) Individual differences in rater decision-making style: An exploratory mixed-methods study, *Language Assessment Quarterly* 9, 225–248.

Banerjee, H L (2019) Investigating the construct of topical knowledge in second language assessment: A scenario-based assessment approach, *Language Assessment Quarterly* 16, 133–160.

Barkaoui, K (2007) Rating scale impact on EFL essay marking: A mixed-method study, *Assessing Writing* 12, 86–107.

Barkaoui, K (2010) Variability in ESL essay rating processes: The role of the rating scale and rater experience, *Language Assessment Quarterly* 7, 54–74.

Barkaoui, K, Brooks, L, Swain, M and Lapkin, S (2012) Test-takers' strategic behaviors in independent and integrated speaking tasks, *Applied Linguistics* 34, 304–324.

Bartlett, F C (1932) *Remembering: A Study in Experimental and Social Psychology*, Cambridge: Cambridge University Press.

Bei, X (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.

Berry, V, Nakatsuhara, F, Inoue, C and Galaczi, E D (2018) *Exploring the use of video-conferencing technology to deliver the IELTS Speaking Test: Phase 3 technical trial*, IELTS Partnership Research Papers 2018/1, IELTS Partners: British Council, Cambridge Assessment English/IDP: IELTS Australia.

Bond, T and Fox, C (2007) *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, New Jersey: Lawrence Erlbaum.

Brooks, L (2003) Converting an observation checklist for use with the IELTS Speaking Test, *Research Notes* 11, 20–21.

Brown, A (2006) *An examination of the rating process in the revised IELTS Speaking Test*, IELTS Research Reports Volume 6, available online: www.ielts.org/-/media/research-reports/ielts_rr_volume06_report2.ashx.

Brown, A and Hill, K (1998) *Interviewer style and candidate performance in the IELTS oral interview*, IELTS Research Reports 1998 Volume 1, available online: www.ielts.org/-/media/research-reports/ielts_rr_volume01_report1.ashx.

Brown, A and Taylor, L (2006) A worldwide survey of examiners' views and experience of the revised IELTS Speaking Test, *Research Notes* 26, 14–18.

Brown, A, Iwashita, N and McNamara, T (2005) *An Examination of Rater Orientations and Test-Taker Performance on English-for-Academic-Purposes Speaking Tasks*, ETS TOEFL Monograph Series MS-29, available online: www.ets.org/Media/Research/pdf/RR-05-05.pdf.

Brown, G, Anderson, A, Shillcock, R and Yule, G (1985) *Teaching Talk: Strategies for Production and Assessment*, Cambridge: Cambridge University Press.

Bui, G (2014) Task readiness: Theoretical framework and empirical evidence from topic familiarity, strategic planning and proficiency levels, *Processing Perspectives on Task Performance* 5, 63–93.

Bui, G and Huang, Z (2018) L2 fluency as influenced by content familiarity and planning: Performance, measurement, and pedagogy, *Language Teaching Research* 22, 94–114.

Cai, Y and Kunnan, A J (2019) Detecting the language thresholds of the effect of background knowledge on a Language for Specific Purposes reading performance: A case of the island ridge curve, *Journal of English for Academic Purposes* 42, 100795.

Carrell, P L (1981) Culture-specific schemata in L2 comprehension, in Orem, R A and Haskell, J F (Eds) *Selected papers from the ninth Illinois TESOL/BE annual convention and the first Midwest TESOL Conference*, Chicago: TESOL/BE, 123–132.

Carrell, P L and Eisterhold, J C (1983) Schema theory and ESL reading pedagogy, *TESOL Quarterly* 17, 553–573.

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* 43 (2), 285–309.

Chapelle, C A and Chung, Y-R (2010) The promise of NLP and speech processing technologies in language assessment, *Language Testing* 27, 301–315.

Chen, L, Zechner, K, Yoon, S, Evanini, K, Wang, X, Loukina, A, Tao, J, Davis, L, Lee, C M and Ma, M (2018) *Automated scoring of nonnative speech using the speechrater sm v. 5.0 engine*, ETS Research Report Series 2018/1, available online: onlinelibrary.wiley.com/doi/full/10.1002/ets2.12198.

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, 209–227.

Cheung, K Y F, McElwee, S and Emery, J (Eds) (2017) *Applying the Socio-cognitive Framework to the BioMedical Admissions Test: Insights from Language Assessment*, Studies in Language Testing volume 49, Cambridge: UCLES/Cambridge University Press.

Chiang, C S and Dunkel, P (1992) The effect of speech modification, prior knowledge, and listening proficiency on EFL lecture learning, *TESOL Quarterly* 26, 345–374.

Cizek, G J (2011) *Reconceptualizing validity and the place of consequences*, paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA, 9–11 April 2011.

Clapham, C (1996) *The Development of IELTS: A Study of the Effect of Background on Reading Comprehension*, Studies in Language Testing volume 4, Cambridge: UCLES/Cambridge University Press.

Clapham, C (2000) Assessment for academic purposes: where next?, *System* 28, 511–521.

Cohen, A D, Segal, M and Bar-Siman-To, R (1984) The C-test in Hebrew, *Language Testing* 1, 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.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.

Creswell, J W and Plano Clark, V (2007) *Designing and Conducting Mixed Methods Research*, Thousand Oaks: Sage.

Creswell, J W and Creswell, J D (2017) *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, Thousand Oaks: Sage.

Cronbach, L J (1949) *Essentials of Psychological Testing*, New York: Harper.

Cronbach, L J (1951) Coefficient alpha and the internal structure of tests, *Psychometrika* 16, 297–334.

Cronbach, L J (1988) Five perspectives on the validity argument, in Wainer, H and Braun, H I (Eds) *Test Validity*, New Jersey: Lawrence Erlbaum, 3–17.

Cronbach, L J (1990) *Essentials of Psychological Testing*, New York: Harper & Row.

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, 67–96.

Davies, A (2008) *Assessing Academic English: Testing English Proficiency, 1950–1989: The IELTS Solution*, Studies in Language Testing volume 23, Cambridge: UCLES/Cambridge University Press.

De Bot, K (1992) A bilingual production model: Levelt's 'Speaking' model adapted, *Applied Linguistics* 13, 1–24.

De Jong, N H (2018) Fluency in second language testing: Insights from different disciplines, *Language Assessment Quarterly* 15, 237–254.

De Jong, N H and Vercellotti, M L (2016) Similar prompts may not be similar in the performance they elicit: Examining fluency, complexity, accuracy, and lexis in narratives from five picture prompts, *Language Teaching Research* 20, 387–404.

DiPrete, T A and Eirich, G M (2006) Cumulative advantage as a mechanism for inequality: A review of theoretical and empirical developments, *Annual Review of Sociology* 32, 271–297.

Dorans, N J and Feigenbaum, M D (1994) *Technical Issues Related to the Introduction of the new SAT and PSAT/NMSQT*, ETS Research Memorandum, RM-94-10, Princeton: Educational Testing Service.

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 Series 2009, Princeton: Educational Testing Service.

Douglas, D and Selinker, L (1992) Analyzing oral proficiency test performance in general and specific purpose contexts, *System* 20, 317–328.

Ducasse, A M and Brown, A (2011) *The role of interactive communication in IELTS Speaking and its relationship to candidates' preparedness for study or training contexts*, IELTS Research Reports Volume 12, available online: www.ielts.org/-/media/research-reports/ielts-rr-volume-12-report-3.ashx.

Eckes, T (2009) Many-facet Rasch measurement, in Takala, S (Ed) *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H), Strasbourg: Council of Europe/Language Policy Division.

Eckes, T (2019) *Many-facet Rasch Measurement: Implications for Rater-mediated Language Assessment*, Quantitative Data Analysis for Language Assessment Volume I: Fundamental Techniques, Abingdon: Routledge.

Eckes, T and Grotjahn, R (2006) A closer look at the construct validity of C-tests, *Language Testing* 23, 290–325.

Edgeworth, F Y (1890) The element of chance in competitive examinations, *Journal of the Royal Statistical Society* 53, 644–663.

Elder, C, McNamara, T F, Kim, H, Pill, J and Sato, T (2017) Interrogating the construct of communicative competence in language assessment contexts:

What the non-language specialist can tell us, *Language & Communication* 57, 14–21.

Ellis, R and Barkhuizen, G (2005) *Analysing Learner Language*, Oxford: Oxford University Press.

Fanselow, J F (1977) Beyond Rashomon: conceptualizing and describing the teaching act, *TESOL Quarterly* 11 (1), 17–39.

Feldmann, U and Stemmer, B (1987) Thin_ aloud a_ retrospective da_ in C-te_ taking: diffe_ languages-diff_ learners-sa_ approaches, in Faerch, C and Kasper, G (Eds) *Introspection in Second Language Research*, Philadelphia: Multilingual Matters, 251–267.

Field, A (2010) *Discovering Statistics using SPSS (and sex and drugs and rock'n'roll)* (Third edition), London: Sage.

Field, J (2011) Cognitive validity, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge UCLES/Cambridge University Press, 65–111.

Fulcher, G (1993) *The construction and validation of rating scales for oral tests in English as a foreign language*, PhD thesis, University of Lancaster.

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

Fulcher, G and Davidson, F (2007) *Language Testing and Assessment*, Abingdon: Routledge.

Fulcher, G and Márquez Reiter, R (2003) Task difficulty in speaking tests, *Language Testing* 20, 321–344.

Geranpayeh, A and Taylor, L (Eds) (2013) *Examining Listening: Research and Practice in Assessing Second Language Listening*, Studies in Language Testing volume 35, Cambridge: UCLES/Cambridge University Press.

Goetz, J P and Le Compte, M D (1984) *Ethnography and Qualitative Design in Educational Research*, Orlando: Academic Press.

Hambleton, R and Rodgers, J (1994) Item bias review, *Practical Assessment, Research, and Evaluation* 4 (6), available online: scholarworks.umass.edu/pare/vol4/iss1/6.

Hastings, A J (2002) Error analysis of an English C-Test: evidence for integrated processing, in Grotjahn, R (Ed) *Der C-Test: theoretische Grundlagen und praktische Anwendungen* [*The C-test: Theoretical Foundations and Practical Applications*] *Volume 4*, Bochum: AKS-Verlag, 53–66.

He, L and Shi, L (2012) Topical knowledge and ESL writing, *Language Testing* 29, 443–464.

Heritage, J (2012) The epistemic engine: Sequence organization and territories of knowledge, *Research on Language & Social Interaction* 45, 30–52.

Herring, S C (2018) The co-evolution of computer-mediated communication and computer-mediated discourse analysis, in Bou-Franch, P and Garcés-Conejos Blitvich, P (Eds) *Analysing Digital Discourse: New Insights and Future Directions*, London: Palgrave Macmillan, 25–67.

Huang, H-T D (2010) *Modeling the relationships among topical knowledge, anxiety, and integrated speaking test performance: a structural equation modeling approach*, dissertation, University of Texas.

Huang, H-T D, Hung, S-T A and Plakans, L (2018) Topical knowledge in L2 speaking assessment: Comparing independent and integrated speaking test tasks, *Language Testing* 35, 27–49.

IBM (2010) *SPSS 19.0*, Armonk: IBM Corp.

Inoue, C (2013) *Investigating the use of language functions for validating speaking test specifications*, paper presented at the Language Testing Forum, Nottingham, UK.

Inoue, C, Khabbazbashi, N, Lam, D and Nakatsuhara, F (2021) *Towards New Avenues for the IELTS Speaking Test: Insights from Examiners' Voices*, IELTS Research Reports Online Series No 2, British Council/Cambridge Assessment English/IDP: IELTS Australia.

Isaacs, T (2018) Shifting sands in second language pronunciation teaching and assessment research and practice*, Language Assessment Quarterly* 15, 273–293.

Jamieson, J M, Eignor, D, Grabe, W and Kunnan, A J (2008) *Frameworks for a new TOEFL: Building a Validity Argument for the Test of English as a Foreign Language*, New York: Routledge.

Jennings, M, Fox, J, Graves, B and Shohamy, E G (1999) The test-takers' choice: an investigation of the effect of topic on language-test performance, *Language Testing* 16, 426–456.

Jensen, C and Hansen, C (1995) The effect of prior knowledge on EAP listening-test performance, *Language Testing* 12, 99–119.

Johnson, P (1982) Effects on reading comprehension of building background knowledge, *TESOL Quarterly* 16, 503–516.

Kamimoto, T (1992) An inquiry into what a C-test measures, *Fukuoka Women's Junior College Studies* 44, 67–79.

Kamir, O (2000) Judgment by film: Socio-legal functions of Rashomon, *Yale Journal of Law & Humanities* 12, 39–50.

Kane, M T (2001) Current concerns in validity theory, *Journal of Educational Measurement* 38, 319–342.

Kasper, G and Ross, S J (2007) Multiple questions in oral proficiency interviews, *Journal of Pragmatics* 39, 2,045–2,070.

Kenyon, D M and Malabonga, V (2001) Comparing examinee attitudes toward computer-assisted and other proficiency assessments, *Language Learning & Technology* 5, 60–83.

Kenyon, D M and Malone, M (2010) Investigating examinee autonomy in a computerized test of oral proficiency, in Araújo, L (Ed) *Computer-based Assessment of Foreign Language Speaking Skills*, Luxembourg: Publications Office of the European Union, 1–27.

Khabbazbashi, N and Galaczi, E D (2020) A comparison of holistic, analytic, and part marking models in speaking assessment, *Language Testing* 37, 333–360.

Khalifa, H and Salamoura, A (2011) Criterion-related validity, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press, 259–292.

Khalifa, H and Weir, C J (2009) *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing volume 29, Cambridge: UCLES/Cambridge University Press.

Kim, Y-H (2009) An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach, *Language Testing* 26, 187–217.

Kirk, R E (1996) Practical significance: A concept whose time has come, *Educational and Psychological Measurement* 56, 746–759.

Klein-Braley, C (1997) C-Tests in the context of reduced redundancy testing: An appraisal, *Language Testing* 14, 47–84.

Klein-Braley, C and Raatz, U (1984) A survey of research on the C-Test1, *Language Testing* 1, 134–146.

Krekeler, C (2006) Language for special academic purposes (LSAP) testing: the effect of background knowledge revisited, *Language Testing* 23, 99–130.

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 Modeling Approach*, Studies in Language Testing volume 2, Cambridge: UCLES/Cambridge University Press.

Lee, H-K and Anderson, C (2007) Validity and topic generality of a writing performance test, *Language Testing* 24, 307–330.

Lee, Y-W (2006) Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks, *Language Testing* 23, 131–166.

Lee-Ellis, S (2009) The development and validation of a Korean C-Test using Rasch Analysis, *Language Testing* 26, 245–274.

Levelt, W J (1989) *Speaking: From Intention to Articulation*, Cambridge: MIT Press.

Li, C-H, Chen, C-J, Wu, M-J, Kuo, Y-C, Tseng, Y-T, Tsai, S-Y and Shih, H-C (2017) The effects of cultural familiarity and question preview type on the listening comprehension of L2 learners at the secondary level, *International Journal of Listening* 31, 98–112.

Lim, G S (2009) *Prompt and Rater Effects in Second Language Writing Performance Assessment*, PhD thesis, University of Michigan.

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, 543–560.

Linacre, J M (1989) *Many-facet Rasch Measurement*, Chicago: MESA Press.

Linacre, J M (2018a) *A User's Guide to FACETS: Rasch Model Computer Programs*, available online: www.winsteps.com.

Linacre, J M (2018b) *Facets Rasch measurement computer program*, version 3.81, available online: www.winsteps.com.

Long, D R (1990) What you don't know can't help you: An exploratory study of background knowledge and second language listening comprehension, *Studies in Second Language Acquisition* 12 (1), 65–80.

Luce, R D and Tukey, J W (1964) Simultaneous conjoint measurement: A new type of fundamental measurement, *Journal of Mathematical Psychology* 1, 1–27.

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

Lumley, T and McNamara, T F (1995) Rater characteristics and rater bias: Implications for training, *Language Testing* 12, 54–71.

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, 415–437.

Luoma, S (2004) *Assessing Speaking*, Cambridge: Cambridge 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, 205–223.

Marais, I and Andrich, D (2008) Formalizing dimension and response violations of local independence in the unidimensional Rasch model, *Journal of Applied Measurement* 9, 200–215.

Marais, I and Andrich, D (2011) Diagnosing a common rater halo effect using the polytomous Rasch model, *Journal of Applied Measurement* 12, 194–211.

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, 157–170.

Masters, G N (1982) A Rasch model for partial credit scoring, *Psychometrika* 47, 149–174.

McNamara, T F (1996) *Measuring Second Language Performance*, Boston: Addison Wesley Longman.

McNamara, T F (2000) *Language Testing*, Oxford: Oxford University Press.

McNamara, T F and Roever, C (2006) *Language Testing: The Social Dimension*, Oxford: Blackwell Publishing.

McNamara, T F, Knoch, U and Fan, J (2019) *Fairness, Justice & Language Assessment*, Oxford: Oxford University Press.

Messick, S (1989), Validity, in Linn, R L (Ed) *Educational Measurement* (Third edition), New York: American Council on Education/Macmillan, 13–103.

Messick, S (1996) Validity and washback in language testing, *Language Testing* 13, 241–256.

Milanovic, M and Saville, N (Eds) (1996) *Performance Testing, Cognition and Assessment: Selected Papers from the 15th Language Research Testing Colloquium, Cambridge and Arnhem*, Studies in Language Testing volume 3, Cambridge: UCLES/Cambridge University Press.

Milanovic, M, Saville, N and Shuhong, S (1996) A study of the decision-making behaviour of composition markers, in Milanovic, M and Saville, N (Eds) *Performance Testing, Cognition and Assessment: Selected Papers from the 15th Language Research Testing Colloquium, Cambridge and Arnhem*, Studies in Language Testing volume 3, Cambridge: UCLES/Cambridge University Press, 92–114.

Moeller, A J, Creswell, J W and Saville, N (Eds) (2016) *Second Language Assessment and Mixed Methods Research*, Studies in Language Testing volume 43, Cambridge: UCLES/Cambridge University Press.

Myford, C M and Wolfe, E W (2000) *Monitoring Sources of Variability within the Test of Spoken English Assessment System*, ETS Research Report 65, June 2000, Princeton: New Jersey.

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, 386–422.

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, 189–227.

Nakatsuhara, F (2011) Effects of test-taker characteristics and the number of participants in group oral tests, *Language Testing* 28, 483–508.

Nakatsuhara, F (2014) *A research report on the development of the Test of English for Academic Purposes (TEAP) speaking test for Japanese university entrants – Study 1 & Study 2*, available online: www.eiken.or.jp/teap/group/pdf/teap_speaking_report1.pdf.

Nakatsuhara, F (2018) Rational design: The development of the IELTS Speaking test, in Seedhouse, P and Nakatsuhara, F, *The Discourse of the IELTS Speaking Test: Interactional Design and Practice*, English Profile Studies 7, Cambridge: UCLES/Cambridge University Press, 17–44.

Nakatsuhara, F, Inoue, C and Taylor, L (2017) *Investigation into double-marking methods: comparing live, audio and video rating of performance on the IELTS Speaking Test*, available online: www.ielts.org/-/media/research-reports/ielts_online_rr_2017-1.ashx.

Nakatsuhara, F, Inoue, C, Berry, V and Galaczi, E D (2017) Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study, *Language Assessment Quarterly* 14, 1–18.

Nitta, R and Nakatsuhara, F (2014) A multifaceted approach to investigating pre-task planning effects on paired oral test performance, *Language Testing* 31, 147–175.

Norris, J M (2016) Current uses for task-based language assessment, *Annual Review of Applied Linguistics* 36, 230–244.

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, 395–418.

Norris, J M, Brown, J D, Hudson, T D and Yoshioka, J (1998) *Designing Second Language Performance Assessments, Technical Report*, Hawaii: University of Hawaii Press.

O'Grady, S (2019) The impact of pre-task planning on speaking test performance for English-medium university admission, *Language Testing* 36, 505–526.

O'Reilly, T and Sabatini, J (2013) *Reading for Understanding: How Performance Moderators and Scenarios Impact Assessment Design*, ETS Research Report Series RR-13-31, Princeton: Educational Testing Service.

Ortega, L (1999) Planning and focus on form in L2 oral performance, *Studies in Second Language Acquisition* 21 (1), 109–148.

O'Sullivan, B (2000) *Towards a model of performance in oral language testing*, PhD thesis, University of Reading.

O'Sullivan, B (2002) Learner acquaintanceship and oral proficiency test pair-task performance, *Language Testing* 19, 277–295.

O'Sullivan, B (2011) *Language Testing: Theories and Practices*, Basingstoke: Palgrave Macmillan.

O'Sullivan, B (2016) Adapting tests to the local context, *New Directions in Language Assessment: JASELE journal*, 145–158.

O'Sullivan, B and Green, A (2011) Test taker characteristics, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press, 36–64.

O'Sullivan, B and Nakatsuhara, F (2011) Quantifying conversational styles in group oral test discourse, in O'Sullivan, B (Ed) *Language Testing: Theories and Practices*, Basingstoke: Palgrave Macmillan, 164–185.

O'Sullivan, B and Rignall, M (2007) Assessing the value of bias analysis feedback to raters for the IELTS writing module, in Taylor, L and Falvey, P (Eds) *IELTS Collected Papers: Research in Speaking and Writing Assessment*, Studies in Language Testing volume 19, Cambridge: UCLES/Cambridge University Press, 446–478.

O'Sullivan, B and Weir, C J (2011) Language testing and validation, in O'Sullivan, B (Ed) *Language Testing: Theory & Practice*, Oxford: Palgrave, 13–32.

O'Sullivan, B, Weir, C J and Saville, N (2002) Using observation checklists to validate speaking-test tasks, *Language Testing* 19, 33–56.

Pallant, J F and Tennant, A (2007) An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS), *British Journal of Clinical Psychology* 46, 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, 52–81.

Pollitt, A and Murray, N L (1996) What raters *really* pay attention to, in Milanovic, M and Saville, N (Eds) (1996) *Performance Testing, Cognition and Assessment: Selected Papers from the 15th Language Research Testing Colloquium, Cambridge and Arnhem*, Studies in Language Testing volume 3, Cambridge: UCLES/Cambridge University Press, 74–91.

Purpura, J E (2016) Second and foreign language assessment, *The Modern Language Journal* 100, 190–208.

Qiu, X (2020) Functions of oral monologic tasks: Effects of topic familiarity on L2 speaking performance, *Language Teaching Research* 24, 745–764.

Raatz, U and Klein-Braley, C (2002) Introduction to language testing and to C-Tests, *University Language Testing and the C-Test*, 75–91.

Rasch, G (1960) *Studies in Mathematical Psychology I. Probabilistic Models for Some Intelligence and Attainment Tests*, Copenhagen: Nielsen & Lydiche.

Read, J (1990) Providing relevant content in an EAP writing test, *English for Specific Purposes* 9, 109–121.

Révész, A (2014) Towards a fuller assessment of cognitive models of task-based learning: Investigating task-generated cognitive demands and processes, *Applied Linguistics* 35, 87–92.

Robinson, P (2001) Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA, in Robinson, P (Ed) *Cognition and Second Language Instruction*, Cambridge: Cambridge University Press, 287–318.

Rosenbaum, P R (1988) Items bundles, *Psychometrika* 53, 349–359.

Rumelhart, D E (1980) *Schemata: The Building Blocks of Cognition*, Theoretical Issues in Reading Comprehension, New Jersey: Lawrence Erlbaum.

Sacks, H (1992) *Lectures on Conversation Volumes One and Two*, Oxford: Blackwell Publishing.

Sato, T (2012) The contribution of test-takers' speech content to scores on an English oral proficiency test, *Language Testing* 29, 223–241.

Schegloff, E A (2007) *Sequence Organization in Interaction: A Primer in Conversation Analysis I*, Cambridge: Cambridge University Press.

Schmidt-Rinehart, B C (1994) The effects of topic familiarity on second language listening comprehension, *The Modern Language Journal* 78, 179–189.

Seedhouse, P (2018) Topic: A key construct with a dual personality, in Seedhouse, P and Nakatsuhara, F, *The Discourse of the IELTS Speaking Test: Interactional Design and Practice*, English Profile Studies 7, Cambridge: UCLES/Cambridge University Press, 114–158.

Seedhouse, P and Egbert, M (2006) The interactional organisation of the IELTS Speaking test, *IELTS Research Reports* 6, 161–206.

Seedhouse, P and Harris, A (2011) Topic development in the IELTS Speaking test, *IELTS Research Reports* 12, 55–110.

Seedhouse, P and Nakatsuhara, F (2018) *The Discourse of the IELTS Speaking Test: Interactional Design and Practice*, English Profile Studies 7, Cambridge: UCLES/Cambridge University Press.

Shaw, S D and Weir, C J (2007) *Examining Writing: Research and Practice in*

Assessing Second Language Writing, Studies in Language Testing volume 26, Cambridge: UCLES/Cambridge University Press.

Shohamy, E G (2001) *The Power of Tests: A Critical Perspective on the Uses of Language Tests*, London: Pearson Education.

Singleton, D and Singleton, E (2002) The C-Test and L2 acquisition/processing research, *University Language Testing and the C-Test*, 143–168.

Skehan, P (1996) A framework for the implementation of task-based instruction, *Applied Linguistics* 17, 38–62.

Skehan, P (1998) *A Cognitive Approach to Language Learning*, Oxford: Oxford University Press.

Skehan, P (2001) *Task and Language Performance Assessment, Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing*, London: Longman.

Skehan, P and Foster, P (1997) Task type and task processing conditions as influences on foreign language performance, *Language Teaching Research* 1, 185–211.

Skehan, P and Foster, P (1999) The influence of task structure and processing conditions on narrative retellings, *Language Learning* 49, 93–120.

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, 170–187.

Smith, J A (1989) Topic and variation in ITA oral proficiency: SPEAK and field-specific tests, *English for Specific Purposes* 8, 155–167.

Smith, J A (1992) *Topic and variation in the oral proficiency of international teaching assistants*, PhD dissertation, University of Minnesota.

Smith, S (2009) *IELTS examination preparation among University of Oxford post-graduate students*, unpublished MSc thesis, University of Oxford.

Spolsky, B (1981) Some ethical questions about language testing, *Practice and Problems in Language Testing* 1, 5–21.

Strauss, A L and Corbin, J M (1998) *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*, Thousand Oaks: Sage.

Swain, M (1984) Teaching and testing communicatively, *TESL Talk* 15, 7–18.

Taghizadeh Vahed, S and Alavi, S M (2020) The role of discipline-related knowledge and test task objectivity in assessing reading for academic purposes, *Language Assessment Quarterly* 17, 1–17.

Tarone, E (1988) *Variation in Interlanguage*, London: Hodder Arnold.

Tarone, E (1998) Research on interlanguage variation: Implications for language testing, in Bachman, L F and Cohen, A D (Eds) *Interfaces Between Second Language Acquisition and Language Testing Research*, Cambridge: Cambridge University Press, 71–89.

Tashakkori, A and Teddlie, C (2010) *Sage Handbook of Mixed Methods in Social & Behavioral Research*, Thousand Oaks: Sage.

Taylor, L (2007a) The impact of the joint-funded research studies on the IELTS Speaking Module, in Taylor, L and Falvey, P (Eds) *IELTS Collected Papers: Research in Speaking and Writing Assessment*, Studies in Language Testing volume 19, Cambridge: UCLES/Cambridge University Press, 185–194.

Taylor, L (2007b) Introduction, in Taylor, L and Falvey, P (Eds) *IELTS Collected Papers: Research in Speaking and Writing Assessment*, Studies in Language Testing volume 19, Cambridge: UCLES/Cambridge University Press, 1–34.

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

Taylor, L and Galaczi, E D (2011) Scoring validity, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press, 171–233.

Tedick, D J (1990) ESL writing assessment: Subject-matter knowledge and its impact on performance, *English for Specific Purposes* 9, 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, 1,358–1,362.

Time Out (2007) *'Time Out' 1000 Books to Change Your Life*, London: Time Out Group Ltd.

University of Cambridge ESOL Examinations (2002) *Cambridge IELTS 3: Examination papers from the University of Cambridge Local Examinations Syndicate*, Cambridge: Cambridge University Press.

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 (2006a) *Cambridge IELTS 5: Examination papers from University of Cambridge ESOL Examinations*, Cambridge: Cambridge University Press.

University of Cambridge ESOL Examinations (2006b) *IELTS Scores Explained* (DVD), 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 (2009) *Cambridge IELTS 7: Examination papers from University of Cambridge ESOL Examinations*, Cambridge: Cambridge University Press.

Usó-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, 210–227.

Van Wyke, J and Andrich, D (2006) A typology of polytomously scored mathematics items disclosed by the Rasch model: implications for constructing a continuum of achievement, in Andrich, D and Luo, G (Eds) *Report no. 2 ARC linkage grant LP0454080: Maintaining invariant scales in state, national and international assessments*, Perth: Murdoch University.

Vann, R J, Lorenz, F O and Meyer, D M (1984) Error gravity: Faculty response to errors in the written discourse of nonnative speakers of English, *TESOL Quarterly* 18 (3), 427–440.

Wang, Z, Zechner, K and Sun, Y (2018) Monitoring the performance of human and automated scores for spoken responses, *Language Testing* 35, 101–120.

Weigle, S C (1998) Using FACETS to model rater training effects, *Language Testing* 15, 263–287.

Weigle, S C (2004) Integrating reading and writing in a competency test for non-native speakers of English, *Assessing Writing* 9, 27–55.

Weiner, B (1992) *Human Motivation: Metaphors, Theories, and Research*, Newbury Park: Sage.

Weir, C J (2005) *Language Testing and Validation: An Evidence-based Approach*, Basingstoke: Palgrave Macmillan.

Weir, C J and Wu, J (2006) Establishing test form and individual task comparability: A case study of a semi-direct speaking test, *Language Testing* 23, 167–197.

Weir, C J, O'Sullivan, B and Horai, T (2006) *Exploring difficulty in Speaking tasks: An intra-task perspective*, available online: www.ielts.org/-/media/research-reports/ielts_rr_volume06_report5.ashx.

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

Wigglesworth, G (1997) An investigation of planning time and proficiency level on oral test discourse, *Language Testing* 14, 85–106.

Wind, S A (2019) Examining the impacts of rater effects in performance assessments, *Applied Psychological Measurement* 43, 159–171.

Wind, S A and Engelhard Jr, G (2013) How invariant and accurate are domain ratings in writing assessment?, *Assessing Writing* 18, 278–299.

Wiseman, C S (2012) Rater effects: Ego engagement in rater decision-making, *Assessing Writing* 17, 150–173.

Wolfe, E W and McVay, A (2012) Application of latent trait models to identifying substantively interesting raters, *Educational Measurement: Issues and Practice* 31, 31–37.

Wolfe, E W, Jiao, H and Song, T (2015) A family of rater accuracy models, *Journal of Applied Measurement* 16, 153–160.

Wright, B D and Masters, G N (2002) Number of person or item strata, *Rasch Measurement Transactions* 16 (3), 888.

Yan, X (2014) An examination of rater performance on a local oral English proficiency test: A mixed-methods approach, *Language Testing* 31, 501–527.

Yang, W and Kim, Y (2020) The effect of topic familiarity on the complexity, accuracy, and fluency of second language writing, *Applied Linguistics Review* 11, 79–108.

Yorozuya, R and Oller Jr, J W (1980) Oral proficiency scales: Construct validity and the halo effect, *Language Learning* 30, 135–153.

Author index

Subject index