Validating Second Language Reading Examinations

Establishing the validity of the GEPT through alignment with the Common European Framework of Reference

For a complete list of titles please visit: www.cambridge.org/elt/silt

Also in this series:

Issues in Testing Business English: The revision of the Cambridge Business English Certificates
Barry O'Sullivan

European Language Testing in a Global Context: Proceedings of the ALTE Barcelona Conference July 2001
Edited by Cyril J. Weir and Michael Milanovic

IELTS Collected Papers: Research in speaking and writing assessment
Edited by Lynda Taylor and Peter Falvey

Testing the Spoken English of Young Norwegians: A study of testing validity and the role of 'smallwords' in contributing to pupils' fluency
Angela Hasselgreen

Changing Language Teaching through Language Testing: A washback study
Liying Cheng

The Impact of High-stakes Examinations on Classroom Teaching: A case study using insights from testing and innovation theory
Dianne Wall

Assessing Academic English: Testing English proficiency 1950–1989 – the IELTS solution
Alan Davies

Impact Theory and Practice: Studies of the IELTS test and *Progetto Lingue 2000*
Roger Hawkey

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

Multilingualism and Assessment: Achieving transparency, assuring quality, sustaining diversity – Proceedings of the ALTE Berlin Conference, May 2005
Edited by Lynda Taylor and Cyril J. Weir

Examining FCE and CAE: Key issues and recurring themes in developing the First Certificate in English and Certificate in Advanced English exams
Roger Hawkey

Language Testing Matters: Investigating the wider social and educational impact of assessment – Proceedings of the ALTE Cambridge Conference, April 2008
Edited by Lynda Taylor and Cyril J. Weir

Components of L2 Reading: Linguistic and processing factors in the reading test performances of Japanese EFL learners
Toshihiko Shiotsu

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

Exploring Language Frameworks: Proceedings of the ALTE Kraków Conference, July 2011
Edited by Evelina D. Galaczi and Cyril J. Weir

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

The Testing Editorial Department, the Language Training and Testing Center (LTTC), Taipei

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

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/9781107443945

First published 2014

Printed in

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

Library of Congress Cataloguing in Publication data

Wu, Rachel Yi-Fen, author.
Validating second language reading examinations : establishing the validity of the GEPT through alignment with The Common European Framework of Reference / Rachel Yi-fen Wu.
pages cm -- (Studies in language testing ; 41)
Includes bibliographical references and index.
ISBN 978-1-107-44394-5 (paperback)
1. English language--Study and teaching--Foreign speakers. 2. English language--Study and teaching--Foreign countries. 3. English language--Ability testing. 4. Second language acquisition--Ability testing. 5. Language and languages--Ability testing. 6. English language--Study and teaching--Foreign countries. I. Title.

PE1128.A2W8 2014
428.0076--dc23

2014021019

For my parents, Ming-kow and Su-yun C Wu

Contents

Appendices

Acknowledgements

I would like to express my gratitude to many individuals who have supported and encouraged me in bringing this book to print.

I am most grateful to Professor Cyril Weir, my PhD supervisor, for his guidance and stimulating advice during my studies, and for his patience and tremendous support throughout the preparation of this book. I would also like to thank my co-supervisor Professor Tony Green for his constant assistance and constructive comments and my external supervisor Professor Lily I-wen Su for her support and encouragement. In addition, I benefited much from Professor Barry O'Sullivan and Dr Fumiyo Nakatsuhara for their insightful feedback on my thesis at the viva voce.

I wish to extend my appreciation to the Language Training and Testing Center (LTTC) for funding this study and providing the General English Proficiency Test (GEPT) data. In particular, I would like to thank Professor Tien-en Kao, former Executive Director, Professor Hsien-hao Liao, Executive Director, and Dr Jessica Wu, Program Director of the R&D Office for their generous support for this project. I also wish to take the opportunity here to acknowledge many colleagues at the Language Training and Testing Center, Cecilia Liao, May Ma, Shih-shen Huang, Matt Li, Mark Pan and Erin Cheng for their assistance and advice with data collection and analysis.

I would like to acknowledge the assistance of Cambridge English Language Assessment for permission to use selections of their examination materials in this study.

For the production of this volume, I would like to thank Dr Evelina Galaczi for her editorial guidance and John Savage for his detailed work of proofing the manuscripts.

This study would not have been possible without the participation of the English as a Foreign Language (EFL) students and GEPT test takers in Taiwan. I am grateful to each of them for their time and effort.

Lastly, my heartfelt thanks go to my parents and my husband for their unconditional support and understanding throughout these years.

Series Editors' note

Since its inception in 1995, the *Studies in Language Testing* (SiLT) series has published a number of PhDs of quality. One of the core purposes of this series is to support and promote work in the field of language assessment by enabling the language testing community to benefit from research which makes a significant contribution to the field, but which might not otherwise reach publication. PhDs are selected for inclusion in the series in accordance with 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.

The first PhD we published was by Anthony Kunnan on test taker characteristics and test performance (SiLT 2) and the next by James E Purpura on learner strategy use and performance (SiLT 8). Eight other PhD theses have been published to date. Caroline Clapham documented the development of IELTS (International English Language Testing System) and looked in particular at the effect of background knowledge on reading comprehension (SiLT 4), while Anthony Green investigated the impact of the IELTS writing subtest on English for Academic Purpose pedagogy (SiLT 25). Kieran O'Loughlin compared direct and semi-direct tests of speaking (SiLT 13) and Angela Hasselgreen looked at testing the spoken English of young Norwegians (SiLT 20). Dianne Wall and Liying Cheng both investigated aspects of test washback and impact, with Wall studying its effects on the classroom in Sri Lanka (SiLT 22) and Cheng carrying out a study on the classroom in Hong Kong (SiLT 21). Toshihiko Shiotsu examined the component of L2 reading ability in the context of Japanese learners of English (SiLT 32). Most recently (SiLT 39) Lynda Taylor investigated how far testing reading through summary tasks enabled us to get closer to measuring the underlying construct of reading ability more faithfully and comprehensively.

SiLT policy is to publish one PhD for every three or four SiLT volumes and in successfully doing this we have enabled high quality doctoral research to reach a wider audience than would normally be expected. In this volume we continue this important tradition and publish a revised version of Rachel Yi-fen Wu's PhD thesis on *Validating Second Language Reading Examinations*.

The appearance of this volume marks the important extension of the socio-cognitive approach to test validation to yet another important suite of high stakes English language tests - this time to the General English Proficiency Test (GEPT) in Taiwan. The GEPT, first administered in 2000, is a 5-level English as a Foreign Language (EFL) testing system, developed by The Language Training and Testing Center (LTTC), Taiwan. The LTTC was established in 1951 and has been providing an extensive range of foreign language teaching and testing services to meet the needs of language education in Taiwan. Wu details how the Ministry of Education lent its support to the LTTC to develop the GEPT under the policy *Towards A Learning Society* to encourage the study of English by providing accessible attainment targets for English learners and engender beneficial washback on the teaching and learning of English in Taiwan. The GEPT is a level-based testing system, designed in accordance with Taiwan's national education framework. The Elementary level is equivalent to that of a junior high school graduate in Taiwan, the Intermediate to that of a senior high school graduate in Taiwan (the age of junior high school students ranges from 13-15, and from 16-18 for senior high school students); the High-Intermediate to that of a university graduate in Taiwan whose major is not English; the Advanced to that of a graduate of a Taiwanese university whose major is English, or to that of a graduate of an English-speaking country, and the Superior to that of a graduate with native English proficiency. Items and content for each GEPT level are designed to match specific level criteria which include a general level description of the overall English proficiency expected at that level and specific skill-level descriptors for the listening, reading, writing and speaking components.

The GEPT is an innovative examination for our times. It has the distinct advantage of being a new examination developed in the modern era. It was thus able to take on board many of the insights arising from the early 21st century work on the socio-cognitive approach and other modern developments in language testing through its purposeful interaction with language testing experts from around the world including Charles Alderson, Lyle Bachman, Anthony Kunnan, Tim McNamara and Cyril Weir.

The examination has benefited from the highly professional approach of its well-trained and well-qualified staff and from the input of outside professionals through annual research awards and a targeted consultancy programme. To maintain and enhance the quality of the GEPT, numerous validation studies have been conducted, including studies on parallel form reliability (Weir and Wu 2006, Wu and Wu 2012), context and cognitive validity (Chan, Wu and Weir 2014), criterion-related validity (Brunfaut and Harding 2014, LTTC 2005, Weir, Chan and Nakatsuhara 2013; Wu and Wu 2010, Yu and Lin 2014), and scoring validity (Wu and Ma 2013).

To provide information for interpreting scores from different tests, Taiwan's Ministry of Education (MOE) selected the Common European Framework of Reference (CEFR) as an international yardstick to benchmark test results. The CEFR, which divides communicative proficiency into six levels arranged in three bands - Basic User (A1 and A2), Independent User (B1 and B2), and Proficient User (C1 and C2) - is intended to 'provide a common basis for elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc.' (Council of Europe 2001:1) and has been used in Europe and wider afield e.g. Japan, to describe curricular aims and learner attainment, as well as to interpret test performance. The MOE considered that the framework suited its need to set English proficiency targets for EFL learners in Taiwan and establish a common platform for comparisons of standards with foreign language educational systems in other countries. Since 2005, the MOE has required all major English exams administered in Taiwan to be mapped against the CEFR. The LTTC thus followed the procedures proposed by the Manual (Council of Europe 2003) to relate the GEPT to the CEFR levels (Wu and Wu 2010). The results showed that the Elementary, Intermediate, High-Intermediate and Advanced levels of the GEPT reading tests are situated at CEFR A2, B1, B2, and C1 levels, respectively.

The core Cambridge English examinations, developed by Cambridge English Language Assessment, are used in this study as the most suitable external measures to provide evidence of criterion-related validity for the GEPT level tests. They were selected because they are among the most highly respected examinations in the field, which have also generated evidence about the relationships of their examinations to the levels of the CEFR. The core Cambridge English examinations 'already ha[ve] an established connection with the CEFR' (Khalifa, ffrench and Salamoura 2010:98), and are 'among a relatively small number of examination[s]' that have applied all three procedures, i.e. 'Specification of the content and purpose', 'Standardisation of interpretation of CEFR levels', and 'Empirical validation studies', recommended by the Manual (Council of Europe 2003) to link with the CEFR (Taylor and Jones 2006:4). There are five levels of the CEFR represented in the core Cambridge English examinations, i.e. *Cambridge English: Key* (KET; also known as Key English Test), *Cambridge English: Preliminary* (PET; also known as Preliminary English Test), *Cambridge English: First* (FCE; also known as First Certificate in English), *Cambridge English: Advanced* (CAE; also known as Cambridge English Advanced), and *Cambridge English: Proficiency* (CPE; also known as Certificate of Proficiency in English). These five tests correspond to CEFR A2, B1, B2, C1, and C2 levels, respectively.

Wu's starting point was that a systematic comparison of the GEPT and core Cambridge English examinations would offer a more grounded specification of proficiency levels at CEFR Levels B1 and B2 than is currently available and in so doing elaborate an efficient methodology for such comparisons that other examination boards might find useful. It would also provide the LTTC and Cambridge with criterion-related validity evidence regarding the constructs underlying their English language assessments at these levels.

As with the LTTC, an overt concern with the constructs being measured by Cambridge English examinations and their relationship to real-life language use was apparent by the beginning of the 21st century. A commitment to transparency and the explicit specification of the communicative content of its examinations was further enhanced by Cambridge's adoption of a socio-cognitive approach to language test design and validation from 2004 onwards; such an approach acknowledged that language use constitutes both a socially situated and a cognitively processed phenomenon and that this must be reflected in language assessment theory and practice.

The increased attention paid to cognitive validity at Cambridge came about as a result of a 10-year project (2004-2013) which saw the publication of the 'construct-focused' volumes in the SiLT series (SiLT 26 (Shaw and Weir 2007), SiLT 29 (Khalifa and Weir 2009), SiLT 30 (Taylor (Ed) 2011) and SiLT 35 (Geranpayeh and Taylor (Eds) 2013)), guided by Mike Milanovic, Nick Saville, Lynda Taylor, Evelina Galaczi and Cyril Weir on the editorial steering committee. This ambitious project enabled far greater attention to be paid than previously to the cognitive processing typically activated in test and non-test tasks, and to the importance of an appropriate match between the two. There is now a widespread acceptance within Cambridge English Language Assessment and its partners, and in the wider international testing community, of the importance for any successful assessment system of seeking and assembling validity evidence on each of the three core aspects of validity: cognitive, context and scoring, which together constitute test construct validity.

Rachel Wu's PhD thesis falls squarely within this paradigm and seeks to ground an empirical framework for test validation and comparison of level-based test batteries and to identify parameters that are useful to explicitly describe different levels of reading proficiency examinations based on a critical evaluation of alignment of the examinations with the CEFR. The scope of the study is limited to CEFR B1 and B2 levels. It uses Weir's (2005) socio-cognitive validation framework, as expanded and more fully explicated for reading in Khalifa and Weir (2009), to establish various aspects of the validity of different levels of the GEPT reading examinations in terms of contextual parameters, cognitive processing skills, and test results. The CEFR and two levels of a CEFR-aligned multilevel test battery, PET and FCE developed by Cambridge ESOL (the former name of Cambridge English Language

Assessment), served as external referents for a review of the similarities and differences between GEPT reading tests targeting CEFR B1 and B2 levels.

The main research questions that this study addresses are:

Research Question 1: Is a GEPT reading test designed to measure at CEFR B2 level more difficult than a GEPT reading test designed to measure at CEFR B1 level in terms of test results, contextual parameters, and cognitive processing skills?

Research Question 2: Are GEPT reading tests at CEFR B1 and B2 levels comparable to alternative CEFR-linked measures in terms of test results, contextual parameters, and cognitive processing skills?

Chapter 1 presents the background to the study, the research objectives, and the research questions. Chapter 2 provides a broad review of the literature on vertical scaling, horizontal comparison of test scores on different tests at an equivalent level, content-based approaches to defining and comparing proficiency levels, and test comparability. To establish the parameters that different language exams adopt to define levels of proficiency, the literature on CEFR alignment, CEFR linking studies, live language proficiency scales which have gained widespread recognition, contextual conditions affecting reading performance, and cognitive processing in reading are surveyed. The literature survey on CEFR alignment covers alignment procedures and CEFR linking studies. The chapter concludes with a discussion of the issues involved in comparing examinations.

In Chapter 3, Wu discusses the research methodology adopted in the study. To answer Research Question 1, the research design and procedures for vertically linking scores from different test levels onto a common score scale are evaluated. To answer Research Question 2, empirical procedures for comparing two different reading tests targeting the same proficiency level are detailed. In addition, qualitative and quantitative procedures available to analyse the contextual features and cognitive operations involved in test performance are presented to answer both Research Questions 1 and 2.

In Chapter 4 the author addresses Research Question 1 by presenting the results of the validation of the tests in the GEPT level framework in terms of test difficulty. Results from vertically linking different levels of the GEPT onto a common score scale are presented, and qualitative and quantitative analyses of contextual parameters and cognitive processing levels are discussed.

In Chapter 5, the author looks at the data generated by the empirical comparison of two different CEFR-aligned English language tests at B1 and B2 levels to answer Research Question 2; scores from the GEPT and Cambridge English reading tests at these levels are presented and comparison of contextual and cognitive parameters in each pair are made.

In Chapter 6, her findings are summarised and the implications of her study for test theory, for test design, for CEFR alignment procedures, and for teaching and course designers are critically discussed. The limitations of the study are then considered, and suggestions for future research put forward.

This volume offers examination boards as well the test developers in the classroom both a practical methodology and the background theoretical support for validating tests of reading comprehension at different proficiency levels. In so doing, it affords them the possibility of their tests laying greater claims to the mantles of cognitive and contextual validity. It provides the means of establishing comparability of reading tests at the same level over time. More importantly, it provides users with a principled basis to empirically establish reading proficiency at the different CEFR levels for their own examinations and ensure criterial differences between levels are operationalised. Perhaps most important of all, it provides the means for different examination boards to compare their examinations with those at a similar CEFR level offered by other examination boards. Similarities will better ground the levels of the examinations offered by examination boards in terms of contextual and cognitive parameters and statistically derived difficulty estimates. Differences should alert the board to the need to justify their interpretation of level in the way they have operationalised these parameters where an examination at a particular level is clearly at odds with other equivalent examinations. Alderson's pithy question: '. . . is my B1 the same as your B1?' can now be addressed.

To make progress in language proficiency testing, it is up to examination boards to co-operate in the same way Cambridge English Language Assessment and LTTC in Taiwan have worked together in this study to compare their English language examinations. Through this volume, the methodology is available for all examination boards offering English language examinations to carry out similar studies. Only in this way will we ever approach a consensus on what an A1 or an A2, a B1 or a B2, a C1 or a C2 level examination in English actually represents/should represent in terms of salient cognitive or contextual parameters and difficulty levels.

Cyril J Weir and Michael Milanovic
June 2014

References

Brunfaut T and Harding, L (2014) *Linking the GEPT Listening Test to the Common European Framework of Reference, LTTC-GEPT Research Report RG-05*. Taipei: Language Training and Testing Center.

Chan, S H C, Wu, R Y F and Weir, C J (2014) *Examining the context and cognitive validity of the GEPT Advanced Writing Task 1: A comparison with real-life academic writing tasks, LTTC-GEPT Research Report RG-03*, Taipei: Language Training and Testing Center.

Council of Europe (2001) *Common European Framework of Reference for*

Languages: Learning, Teaching, Assessment, Cambridge: Cambridge University Press.

Council of Europe (2003) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment: A Manual, Preliminary Pilot Version*, Strasbourg: Council of Europe.

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.

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.

Language Training and Testing Center (2005) *Mapping the GEPT to the Common English Yardstick for English Education in Taiwan, LTTC Research Report*, Taipei: Language Training and Testing Center.

Shaw, S and Weir, C J (2007) *Examining Second Language Writing: Research and Practice in Assessing Second Language Writing,* Studies in Language Testing volume 26, Cambridge: UCLES/Cambridge University Press.

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 Jones, N (2006) ESOL exams and the Common European Framework of Reference (CEFR), *Research Notes* 24, 2–5.

Weir, C J (2005) *Language Testing and Validation: An Evidence-Based Approach*, Hampshire: 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, Chan, S H C and Nakatsuhara, F (2013) *Examining the Criterion-Related Validity of the GEPT Advanced Reading and Writing Tests: Comparing GEPT with IELTS and Real-Life Academic Performance, LTTC-GEPT Research Report RG-01*, Taipei: Language Training and Testing Center.

Wu, J and Ma, T (2013) *Investigating rating processes in an EAP writing test: Insights into scoring validity*, paper presented at the 35th Annual Language Testing Research Colloquium (LTRC), Seoul, Korea.

Wu, J R W and Wu, R Y F (2010) Relating the GEPT reading comprehension tests to the CEFR, in Martyniuk, W (Ed), *Aligning Tests with the CEFR: Reflections on using the Council of Europe's draft Manual,* Studies in Language Testing volume 33, Cambridge: UCLES/Cambridge University Press, 204-223.

Wu, J and Wu, R Y F (2012) *Parallel-Forms Reliability of the GEPT Speaking Test*, paper presented at the 14th Academic Forum of English Language Testing in Asia, Xian, China.

Yu, G and Lin, S (2014) *A Comparability Study on the Cognitive Processes of Taking GEPT (Advanced) and IELTS (Academic) Writing Tasks Using Graph Prompts, LTTC-GEPT Research Report RG-02*. Taipei: Language Training and Testing Center.

Abbreviations

1PL	One-parameter Logistic
2PL	Two-parameter Logistic
3PL	Three-parameter Logistic
ACTFL	American Council on the Teaching of Foreign Languages
ALTE	Association of Language Testers in Europe
ASL	Average Sentence Length
ASLPR	Australian Second Language Proficiency Ratings
ASW	Average Number of Syllables Per Word
AWL	Academic Wordlist
BNC	British National Corpus
CAE	Certificate in Advanced English
CEFR	Common European Framework of Reference for Languages
CLB	Canadian Language Benchmarks
CPE	Certificate of Proficiency in English
CTT	Classical Test Theory
EFL	English as a Foreign Language
ESL	English as a Second Language
ETS	Educational Testing Service
FCE	First Certificate in English
FK	Flesch-Kincaid
FL	Foreign Language
FREs	Flesch Reading Ease Score
FSI	Foreign Service Institute
GEPT	General English Proficiency Test
GESE	Graded Examinations in Spoken English
IRT	Item Response Theory
ISE	Integrated Skills in English
ISLPR	International Second Language Proficiency Ratings
JMLE	Joint Maximum Likelihood Estimation
LSA	Latent Semantic Analysis
LTTC	Language Training and Testing Center
MLE	Maximum Likelihood Estimation
MMLE	Marginal Maximum Likelihood Estimation
MOE	Ministry of Education
NEAT	Non-Equivalent groups Anchor Test
PET	Preliminary English Test

STTR	Standardised Type-token Ratio
TASA	Touchstone Applied Science Associates
TestDaF	Test Deutsch als Fremdsprache
TLU	Target Language Use
TOEFL	Test of English as a Foreign Language
TOEIC	Test of English for International Communication
TTR	Type-token Ratio
UCLES	University of Cambridge Local Examinations Syndicate
WPM	Words Per Minute

1 Introduction

Since the 1970s, with rising public demand for transparent and explicit interpretations of test scores, level-based examinations have received growing attention in the field of language testing. The traditional norm-referenced approach to assessment compares test takers' performance relative to each other without establishing what they are able to do with the language. In contrast, level-based examinations divide language proficiency into defined levels which outline different degrees of achievement and identify whether test takers have attained a criterion standard. Test results are translated into proficiency statements suggesting the language activities that a test taker with a specific score is expected to be able to carry out. The proficiency statements of these level-based examinations are commonly formulated with reference to external standards, such as course objectives, national curricula, or proficiency frameworks that have already gained widespread acceptance to language levels to describe test takers' language competence and to facilitate communication between stakeholders about test objectives.

Recent advances in the fields of applied linguistics and language pedagogy have contributed to the development of numerous language proficiency frameworks in different contexts to reflect 'a hierarchical sequence of performance ranges' (Galloway 1987:27). These proficiency frameworks divide language proficiency into levels that are meaningful to their different users (Brindley 1986, 1991, Richterich and Schneider 1992, Trim 1977). The ones which have gained wide recognition and have continued to be actively used include the International Second Language Proficiency Ratings (ISLPR) Rating; later known as the Australian Second Language Proficiency (ASLPR) Scale (Ingram 1984); the American Council on the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines (Hiple 1987); the Canadian Language Benchmarks (CLB; Pawlikowska-Smith 2000); and the Common European Framework of Reference for Languages (CEFR; Council of Europe 2001).

Among these frameworks, the CEFR has been the most widely used and recognised internationally 'to describe the levels of proficiency required by existing standards, tests and examinations in order to facilitate comparisons between different systems of qualifications' (Council of Europe 2001:21). In the past decade, various language testers and exam boards (e.g. Dunlea and

Matsudaira 2009, Kecker and Eckes 2010, Khalifa, ffrench and Salamoura 2010, Papageorgiou 2007, 2010, Tannenbaum and Wylie 2008, Wu and Wu 2010) followed the procedures that *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment: A Manual, Preliminary Pilot Version* (Council of Europe 2003), commonly referred as the Manual, proposed to align their exams to the CEFR (Council of Europe 2001). They attempted to describe their exams in terms of CEFR levels for the purpose of providing an easily accessible interpretation of test results to their test users and for use in seeking recognition from local governments and international professional organisations.

While the CEFR has been gaining popularity and has contributed to describing test constructs over the past decade, various case studies (e.g. Alderson (Ed) 2002, Figueras and Noijons 2009, Kecker and Eckes 2010, Khalifa et al 2010, Martyniuk and Noijons 2007, Morrow 2004, Wu and Wu 2010) have pointed to the difficulty in using the CEFR to establish proficiency bands in precise terms and call for fuller elaboration of these levels. Westhoff (2007:676) argued that 'although the CEFR descriptors tell us a lot about what learners at a certain level can do, very little is stated about what they should know . . .'. Weir (2005b:12) shared this view and commented that 'the CEFR provides little assistance in identifying the breadth and depth of productive or receptive lexis that might be needed to operate at the various levels.' He argued (2013:434) that examination boards need to 'determine what is an acceptable range for each parameter at each level of proficiency' in order to improve '. . . specifications for the different levels of proficiency which are, at best, vaguely and sparsely specified in the current Common European Frame of Reference.' Alderson, Figueras, Kuijper, Nold, Takala and Tardieu (2006:12) noted that many of the terms in the CEFR are not explicitly defined (e.g. 'long' and 'longer', 'straightforward' and 'complex'), and the CEFR provides no guidance on what structures, lexis or other linguistic features learners might be expected to cope with in order to complete test tasks at various proficiency levels. In addition to the textual features of test tasks, McNamara (1996) and Weir (1993) considered that the cognitive processes engaged by the examinees need to be given equal importance as well so that both the tasks and the conditions under which the tasks are performed can approximate to performance in the real world as closely as possible. In view of the CEFR's inherent limitations, O'Sullivan and Weir (2011) argued that considerable supplementary resources are needed to more comprehensively and explicitly define the levels as described in the CEFR. Weir (2005b:3) proposed that 'a framework is required that helps identify the elements of the context and processing and the relationships between these at varying levels of proficiency, i.e. one that addresses both situational and interactional authenticity (Bachman and Palmer 1996).' To demonstrate the extent of differentiation across exam levels, it will be necessary to identify criterial features of the test tasks and to determine an acceptable range for relative degrees of complexity of each criterial feature at each level of proficiency for which the exam boards offer examinations.

Recognising the need to validate how the constructs of level-based exams may differ according to learners' level of language proficiency, the present study aimed to identify parameters that are useful for developing operationalisable specifications for different levels of reading proficiency and to establish an empirical framework enabling test validation and comparison of examinations developed by different exam boards aiming at the same level. The scope of the study is limited to CEFR B1 and B2 levels. This study applied Weir's (2005a) socio-cognitive validation framework to collect validity evidence of different test levels in terms of contextual parameters, cognitive processing skills, and test results. It focuses on the *cross-level* relationships between two CEFR-aligned reading tests, i.e. the General English Proficiency Test (GEPT) and the core Cambridge English examinations at the B1 and B2 levels.

The GEPT is a 5-level English as a Foreign Language (EFL) testing system, developed by The Language Training and Testing Center (LTTC), Taiwan, in accordance with Taiwan's national education framework. The LTTC, originally named The English Training Center, was established in 1951 to provide training in English for government-sponsored personnel preparing to go to the United States under technical assistance programs in place at that time. In 1986, the Center was registered with the Ministry of Education in Taiwan as a non-profit educational foundation. The LTTC now offers language training and testing in English, Japanese, French, German, and Spanish. In March 1998, the Ministry of Education (MOE) in Taiwan promulgated the *Towards A Learning Society* (邁向學習社會) white paper to promote lifelong learning. Under this policy in 1999, the MOE lent its support to the LTTC to develop the GEPT in order to enhance citizens' motivation for learning English by providing accessible attainment targets for English learners in Taiwan. Test content at the first two levels of the GEPT, i.e. Elementary and Intermediate, is guided by the national curriculum objectives of junior high schools and senior high schools, respectively. The three upper levels of the GEPT, i.e. High-Intermediate, Advanced, and Superior, for which no national curriculum exists, were developed based on the expectations of stakeholders in English education in Taiwan as identified through textbook analysis, needs analysis, and teachers' forums. Items and content for each GEPT level are designed to match specific level criteria which include a general level description of the overall English proficiency expected at that level and specific skill-level descriptors for the listening, reading, writing, and speaking components.

In 2004, the Executive Yuan, the highest administrative body in the government (comparable to the cabinet in other countries), approved 'measures to enhance the English proficiency of civil servants (提升公務人員英語能力改進措施)', a plan undertaken under the policy 'Challenge 2008-National Development Plan (挑戰 2008 國家發展重點計畫)' (2002), and called for 50% of civil servants to pass the GEPT Elementary or Intermediate levels, or other certified equivalent English exams, within three years. To provide information for interpreting scores from different tests, Taiwan's MOE decided to adopt the CEFR as an international yardstick to benchmark test results. The CEFR, which divides communicative proficiency into six levels arranged in three bands – Basic User (A1 and A2), Independent User (B1 and B2), and Proficient User (C1 and C2), is intended to 'provide a common basis for elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc.' (Council of Europe 2001:1) and has been used in Europe and beyond (e.g. Korea and Canada) to describe curricular aims and learner attainment, as well as to interpret test performance; therefore, the Ministry considered that the framework suited its need to set English proficiency targets for EFL learners in Taiwan and establish a common platform for comparisons of standards with foreign language educational systems in other countries. Since 2005, the MOE has required all major English exams administered in Taiwan to be mapped against the CEFR. The LTTC thus followed the procedures proposed by the Manual (Council of Europe 2003) to relate the GEPT to the CEFR levels (Wu and Wu 2010). The results showed that the Elementary, Intermediate, High-Intermediate, and Advanced levels of the GEPT reading tests are situated at CEFR A2, B1, B2, and C1 levels, respectively.

The core Cambridge English examinations, developed by Cambridge English Language Assessment, formerly named University of Cambridge ESOL (English for Speakers of Other Languages) Examinations, were used as external criterion measures to provide evidence of criterion-related validity for the GEPT level tests in this study. They were selected because they are among the few examinations which have made claims about the relationships of their examinations to the levels of the CEFR. The core Cambridge English examinations 'already ha[ve] an established connection with the CEFR' (Khalifa et al 2010:98), and is 'among a relatively small number of examination[s]' that have applied all three procedures, i.e. 'Specification of the content and purpose', 'Standardisation of interpretation of CEFR levels', and 'Empirical validation studies', recommended by the Manual (Council of Europe 2003) to link with the CEFR (Taylor and Jones 2006:4).

The University of Cambridge formed the University of Cambridge Local Examinations Syndicate (UCLES), now Cambridge English Language Assessment, over 150 years ago. Its aims were to raise standards in education by administering exams for people who were not members of the University. Cambridge English Language Assessment provides a variety of examinations covering a wide range of subjects and levels. The five levels of the core Cambridge English examinations are: *Cambridge English: Key* (KET;

also known as Key English Test), *Cambridge English: Preliminary* (PET; also known as Preliminary English Test), *Cambridge English: First* (FCE; also known as First Certificate in English), *Cambridge English: Advanced* (CAE; also known as Certificate in Advanced English), and *Cambridge English: Proficiency* (CPE; also known as Certificate of Proficiency in English). The CPE was first administered in 1913. Following the CPE, UCLES launched the Lower Certificate in English (renamed as FCE in 1975) in 1939, PET in 1980, CAE in 1991 and KET in 1994. These five tests correspond to the Association of Language Testers in Europe (ALTE) Levels 1 to 5 and CEFR A2, B1, B2, C1, and C2 levels, respectively. The five levels reflect the levels of language ability familiar to English language teachers around the world and have been described as 'natural levels' (North 2006:8).

A systematic comparison of the GEPT and the core Cambridge English examinations could potentially provide a more grounded specification of proficiency levels at CEFR B1 and B2 than is currently available and in so doing elaborate an efficient methodology for such comparisons that other examination boards might find useful. It would also provide the LTTC and Cambridge English Language Assessment with validity evidence relating to the constructs underlying their English language assessments at these levels.

The main questions that this study addresses are:

> Research Question 1: Is a GEPT reading test designed to measure at CEFR B2 level more difficult than a GEPT reading test designed to measure at CEFR B1 level in terms of test results, contextual parameters, and cognitive processing skills?
>
> Research Question 2: Are GEPT reading tests at CEFR B1 and B2 levels comparable to alternative CEFR-linked measures in terms of test results, contextual parameters, and cognitive processing skills?

This introductory chapter has provided an outline of the background to the study, the research objectives, and the research questions. Chapter 2 presents a review of related literature on vertical scaling, horizontal comparison of test scores on different tests at an equivalent level, content-based approaches to defining and comparing proficiency levels, and test comparability. A review of vertical scaling includes research on linking different levels of a multilevel exam onto the same vertical scale to provide direction in the construction of data collection and procedures for validation of vertical differentiation of a level-based test, followed by a brief discussion of how scores on a different test at an equivalent level can be used as an external criterion-related check on the validity of a defined level of difficulty. To sort through features that different language exams adopt to define levels of proficiency, the literature on CEFR alignment, CEFR linking studies, language proficiency scales which have gained wide recognition and have continued to be actively used, contextual impacts on reading performance, and cognitive processing in reading, are surveyed. The literature survey on CEFR alignment covers alignment procedures and CEFR linking studies to provide the background to and justification for the present study. This chapter concludes with a discussion of issues involved in comparing examinations.

Chapter 3 discusses the research methodology used in this study. To answer Research Question 1, the research design and procedures for vertically linking scores from different test levels onto a common score scale are described in order to examine whether difficulty increases as the test level advances. To answer Research Question 2, empirical procedures for comparing two different reading tests targeting the same proficiency level are explained to assess whether two reading tests, provided by different exam boards at the same CEFR level, are comparable in terms of test takers' performance. In addition, qualitative and quantitative procedures to analyse contextual features and cognitive operations involved when test takers are responding to the reading tests are presented to answer both Research Questions 1 and 2.

Chapter 4 reports results of the validation of the GEPT level framework in terms of test difficulty, which addresses Research Question 1. Results from vertically linking different levels of the GEPT onto a common score scale are presented, and qualitative and quantitative analyses of contextual features and cognitive processes are discussed.

Chapter 5 reports results from empirical validation comparing two CEFR-aligned tests at the same proficiency level to answer Research Question 2. Results from the empirical comparison between scores from the GEPT and Cambridge English reading tests at CEFR B1 and B2 levels, respectively, are presented. Relationships between test performance and results from qualitative and quantitative analyses of contextual features and cognitive processes are discussed.

Finally, in Chapter 6, a summary of the findings is presented; the implications for test theory, for test design, for CEFR alignment procedures, and for teaching and course designers are discussed. Limitations of the present study are considered, and suggestions for future research are put forward.

2 Defining levels of proficiency: A literature review

Chapter overview

This chapter provides an overview of previous research on test validation in terms of level differentiation. It begins by reviewing vertical scaling studies designed to provide directions for the empirical validation of vertical differentiation across test levels. This is followed by a brief discussion of how scores on a different test at an equivalent level can be used as an external criterion-related check on the validity of the defined level of difficulty. Content-based approaches to specifying proficiency levels suggested in the literature are then reviewed to explore how far they might help an examination board determine the levels of proficiency for its English language tests. Given its widespread use for this purpose, this chapter will focus mainly on the CEFR and the previous linking studies and alignment procedures associated with it. This will provide the background to a growing trend among test developers to use the CEFR to establish test levels, to support the interpretation of test results, and to gain mutual understanding in the field of language education. Some other existing scales, e.g. the ISLPR, the ACTFL Guidelines, and the CLB, that have been used for this purpose, will also be briefly reviewed.

Given the apparent deficiencies in these current approaches to construct definitions, this chapter then turns to recent socio-cognitive approaches to explore what help they can offer examination boards to better define levels of proficiency in their English language examinations. A number of critical components of test specifications will be examined to identify parameters that might be useful for defining levels of language proficiency. This chapter will conclude with discussing issues involved in comparing examinations.

Vertical scaling

Educational tests are scaled or linked statistically for various purposes. Lissitz and Huynh (2003:2) described scaling as a process 'in which raw scores . . . are transformed to a new set of numbers with certain selected attributes, such as a particular mean and standard deviation.' Vertical scaling is to place scores of tests that are 'intentionally designed to be different in difficulty and intended for groups with different ranges of abilities, but which measure the same general area of knowledge or general domain of skills' onto a common score scale (Loyd and Hoover 1980:179). The scale resulting from the process is referred to as 'a vertical scale' or, sometimes, as 'a developmental score scale' (Tong and Kolen 2007:228). Vertical scales provide a systematic way to examine the amounts of developmental change in performance and to investigate differentiation of performance standards across test levels, and thus enrich the interpretations of scores obtained from multilevel test batteries (Patz 2007:22).

Vertical scaling involves challenging psychometric procedures. Creating a vertical scale requires a complicated process of establishing linking relationships so that comparison of scores obtained from tests of systematically different difficulty can be made. Previous research (e.g. Camilli, Yamamoto and Wang 1993, Custer, Omar and Pomplun 2006, Hanson and Beguin 2002, Kim and Cohen 1998, Peterson, Cook and Stocking 1983, Pommerich, Hanson, Harris and Sconing 2004, Tong and Kolen 2007) showed that vertical scaling depended on a variety of factors, such as the linking method, the item response theory (IRT) model, the ability/difficulty estimation method, and the linking procedure used in the construction of the scale. Different choices of linking procedures and scaling methods tend to result in different vertical scales (Camilli et al 1993, Loyd and Hoover 1980, Tong and Kolen 2007, Williams, Pommerich and Tissen 1998, Yen 1986). However, there is no consensus in the literature on which procedure produces the vertical scale that can most adequately capture the nature of growth (Tong and Kolen 2007:228).

The two most commonly used linking procedures for establishing vertical scales are the scaling test design (Peterson, Kolen and Hoover 1989) and the common-item non-equivalent groups design (Kolen and Brennan 2004), also called the Non-Equivalent groups Anchor Test (NEAT) design. The scaling test design requires that a representative sample of examinees from different proficiency levels take both a scaling test and a level test. A scaling test 'is composed of test items that represent the domain of content over all levels of the test . . . and is designed to be administered in a single sitting' (Peterson et al 1989:232), while a level test is designed to be most appropriate for examinees of the respective level. The scaling test is used to place scores from different level tests of the multilevel test battery on the same vertical scale, and the function of level tests is to measure examinees' proficiency.

The common-item non-equivalent groups design requires that examinees take only level tests. Level tests of adjacent levels contain a set of representative common test items and they are administered to different groups of examinees who have systematically different levels of proficiency. Examinees' performances on the items that are common between adjacent levels are used to establish a linking chain to place all levels onto the same vertical scale

(Tong and Kolen 2007:229). The current study employed the common-item non-equivalent groups design, which researchers (Angoff 1971, Kolen and Brennan 2004, Livingston 2004, Peterson et al 1989, Peterson, Marco and Steward 1982) generally favour, due to the practical advantage of not having to compile or to administer an additional scaling test.

To assess whether the experimental design of test forms and the examinee population used to produce vertical scales are adequate, Patz and Yao (2007:253) suggested:

> When differences in population proficiency at adjacent levels are modest in comparison to differences between examinees within levels and when the expectations or standards against which examinees are to be measured overlap extensively, then linking the adjacent test levels to a common scale will make sense and provide meaningful information.

As to the criteria for examining whether the linking procedure is appropriate, Kolen and Brennan (2004:262–263) suggested:

> . . . with the common-item non-equivalent groups design, mean differences between the two groups of approximately .1 or less standard deviation unit on the common items seem to cause few problems for any of the [linking] methods. Mean group differences of around .3 or more standard deviation unit can result in substantial differences among methods, and differences larger than .5 standard deviation unit can be especially troublesome.

To scale scores from multilevel test batteries, the Thurstone method was widely adopted in the 1960s and 1970s, while IRT has received substantial consideration in recent years. Thurstone scaling is based on the assumption that within-level performance is symmetrically distributed, and raw scores (number of items correct) for each group of examinees are converted to normalised *z* scores. However, the assumption of within-level normality is not justifiable in most multilevel educational tests (Peterson et al 1989:236). IRT scaling is based on the assumption that achievement is unidimensional, which means that all items in the test measure a single ability or trait, and models probabilistic distribution of examinees' success at the item level (Yen 1986:302). If data fit the assumption of the IRT model, person abilities can be estimated independent of particular items.

The Thurstone and the IRT methods tend to produce different results. Tong and Kolen (2007:249) observed that '[s]cales developed using the Thurstone method suggested that students grew further apart from each other as they progressed through school years, whereas scales developed using IRT suggested that the spread of students' achievement either fluctuates or decreases over grades.' Furthermore, Yen (1986:300) argued that 'when score scales are not linearly related, different results are produced when calculations are based on one scale rather than the other.'

A variety of IRT models can be used to produce item difficulty and person ability estimates. The most commonly used are the one-parameter (or Rasch), the two-parameter, and the three-parameter logistic (3PL) models. Yen (1986:309–310) pointed out:

> . . . these models differ in their assumptions. If a set of data meets the assumptions of all three models, the models will produce the same scaling; that is, under such circumstances the trait scales produced by the different models will be linearly related. However, if the data are appropriate for the three-parameter model and not the other models, the three methods will produce non-linearly related scales.

The 3PL model (Birnbaum 1968) estimates three characteristics of items:

1. Item discriminating power, i.e. *a* parameter, ranging from 0 to 1.
2. Item difficulty, i.e. *b* parameter, commonly ranging from –3 to 3.
3. Guessing parameter, i.e. *c* parameter, probability that test takers of low ability choose a correct response, ranging from 0 to 1.

The two-parameter logistic (2PL) model estimates only *a* and *b* parameters and sets *c* parameter to 0. The one-parameter logistic (1PL), or Rasch, model, estimates only *b* parameter and sets *c* parameter to 0 and *a* parameter to 1.

Rasch model estimations are generally favoured for linking tests because of their ease of interpretation due to the features of equal interval and item- and person-invariance. These properties do not hold for other IRT models. The Rasch model estimates the log odds probability, named 'logit', a contraction of 'log odds unit'. Logits express relative item difficulties, which are invariant to any specific person, and relative proficiencies of test takers, which are invariant to any specific items. In the Rasch model, there is a one-to-one correspondence between raw scores and logits, and the hierarchical order of test takers by their raw scores and relative distances of raw scores are preserved. In addition to ease of interpretation, another advantage of Rasch model estimations is that a small number of subjects is sufficient to accurately estimate item parameters; Wright and Stone (1979) suggested a sample size of 200 examinees would be enough to perform accurate estimation of the 1PL model. A larger sample size is required to estimate IRT item discriminating parameters than item difficulty parameters (Barnes and Wise 1991), and at least 1,000 (Reckase 1979, Skaggs and Lissitz 1986) to 10,000 (Thissen and Wainer 1982) examinees are needed to accurately estimate the 3PL item parameters. In this study, IRT Rasch model estimations were used to examine relationships between CTT (classical test theory-based) scores on the operational GEPT tests and the IRT item and ability estimates from reading tests at different GEPT levels, and to empirically validate the projected increase in difficulty across test levels.

Horizontal comparison of test scores on different tests at an equivalent level

It is as important to investigate the extent to which scores on a level-based test are comparable with scores on a different measure aiming at an equivalent proficiency level as it is to validate vertical differentiation across test levels. To collect such criterion-related validity evidence, the same group of examinees from the target proficiency level would take the two different tests at approximately the same point in time. Means, standard deviations, overall percentages of items correct, score distributions, a correlation coefficient, and a significance test of differences in means between the two tests targeting the same level are computed to investigate the empirical relationships between the two tests. A correlation coefficient (denoted by *r*) is calculated to examine how strong the relationship is between the scores on the test to be validated, i.e. predictor, and the performance on a different measure, i.e. criterion. If the scores are interval or ratio data, the Pearson Product Moment Correlation is computed. If they are ordinal data, the Spearman Rank Order Correlation (denoted by ρ, pronounced as rho) or the Kendall Rank Order Correlation (denoted by τ, pronounced as tau) is used. A coefficient of –1 or 1 means that there is a perfect, negative or positive respectively, correlation between the two sets of data, while 0 means that there is no linear relationship between them. Once a correlation coefficient is computed, a significance test is normally conducted to determine the probability (denoted by *p*) that the results are due to statistical errors or occur by chance; the smaller the *p* value, the more significant the relationship is between the two sets of data, and the less likely the correlation occurs by chance or due to statistical errors. The most widely-known comparability study of language tests is the 3-year Cambridge-TOEFL Comparability Study (Bachman, Davidson, Ryan and Choi 1995). This study investigated the comparability of the FCE administered by UCLES and the paper-and-pencil version of Test of English as a Foreign Language (TOEFL) administered by the Educational Testing Service (ETS). This research involved a qualitative content analysis of the test tasks and a quantitative analysis of the test performance. The qualitative content analysis of the two tests was conducted by expert judges using the Communicative Language Ability instrument, and the quantitative statistical analysis was conducted by analysing test takers' performances. The results of the study suggested that the two tests generally measured similar language abilities.

Content-based approaches to defining and comparing proficiency levels

A number of content-based approaches to specifying proficiency levels suggested in the literature are discussed below. This section will begin with the CEFR, which has become 'the industry standard' for doing this.

CEFR alignment

The *Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (Council of Europe 2001), developed between 1993 and 1996 by the Council of Europe, has been used in Europe and beyond to describe curricular aims and learner attainment, as well as to interpret test performance. Currently the framework has had a wide-reaching impact on language and education policy worldwide. The major aim of the framework is to provide a common basis for describing language proficiency in order 'to facilitate the mutual recognition of qualifications gained in different learning contexts' (Council of Europe 2001:1) and to 'assist learners, teachers, course designers, examining bodies and educational administrators to situate and co-ordinate their efforts' (2001:6).

The CEFR, which originated from the Council of Europe's Modern Languages Projects in the 1970s, divides communicative proficiency into six levels, arranged in three bands – Basic User (A1 Breakthrough and A2 Waystage), Independent User (B1 Threshold and B2 Vantage), Proficient User (C1 Effective Operational Proficiency and C2 Mastery). The six levels provide convenient points of reference for stakeholders to describe learners' stages of language development. Education professionals are encouraged to merge or subdivide the levels based on the needs specific to their contexts (Council of Europe 2009:3). The CEFR descriptors have been empirically validated on the basis of teachers' perceptions of 'how one might best and consistently describe different levels of actual learner performance' (Byrnes 2007:643). The descriptor scales '[p]rovide a bank of criterion statements about the continuum of foreign language proficiency which can be exploited flexibly for the development of criterion-referenced assessment. They can be matched to existing local systems, elaborated by local experience and/or used to develop new sets of objectives' (Council of Europe 2001:30).

Language testers and exam boards have been making efforts to relate their exams to the CEFR levels for the purpose of providing easily accessible interpretation of test results to their test users and seeking recognition from local governments and international professional organisations in the past decade. Exam boards which have attempted to relate their tests to the CEFR include Test Deutsch als Fremdsprache (TestDaF) (Kecker and Eckes 2010) in Germany; Cambridge English FCE (Khalifa et al 2010), Trinity College

London Graded Examinations in Spoken English (GESE) and Integrated Skills in English (ISE) (Papageorgiou 2007, 2010), and City & Guilds Tests of English (O'Sullivan 2008) in the UK; Test of English for International Communication (TOEIC) and TOEFL (Tannenbaum and Wylie 2008) in the USA; the EIKEN Test (Dunlea and Matsudaira 2009) in Japan; and the GEPT (Wu and Wu 2010) in Taiwan. To support test providers in aligning their examinations to the CEFR levels and validating the linking relationship, the Council of Europe published *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment. A Manual: The Preliminary Pilot Version* in 2003, and following consultation, a revised version was published in 2009. The Manual (Council of Europe 2009) proposes five interrelated sets of procedures, i.e. Familiarisation, Specification, Standardisation Training/Benchmarking, Standard Setting, and Validation, to design a linking scheme and suggests that '[r]elating an examination or test to the CEFR can best be seen as a process of "building an argument" based on a theoretical rationale' (2009:9).

To ensure that the data obtained during the process are of good quality, the Manual (Council of Europe 2003, 2009) provides exercises and materials for those involved in the linking process to familiarise themselves with the CEFR during the Familiarisation procedure. After they have sufficient understanding of the rationale behind the CEFR level, the expert judges follow the rest of the procedures in the Manual to relate the exam(s) to the CEFR. During the Specification procedure, test providers select forms that are relevant to their context from Forms A1 to A23 in the Manual to reflect whether their tests have been developed and administered carefully and followed good practice, as well as to profile the coverage of the examination in relation to the categories presented in the CEFR (Council of Europe 2001) Chapter 4 (Language Use and the Language Learner) and Chapter 5 (The User/learner's Competences) in order to relate their examinations to the CEFR levels.

The Standard Setting procedure involves expert judges to apply their knowledge and experience to reflect their understandings of learners' performance and reach meaningful and relevant judgments on the level of performance required to set cut scores. Cizek (Ed) (2001:5) considers that 'standard setting is perhaps the branch of psychometrics that blends more artistic, political, and cultural ingredients into the mix of its products than any other.' Similarly, the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association and National Council on Measurement in Education 1999:54) suggests that setting cut scores 'embody[s] value judgments as well as technical and empirical considerations.' In light of the challenge for expert judges to develop a common understanding of levels and maintain consistent judgment, the Manual proposes three phases of training, i.e. Phase I Illustration, Phase II

Controlled Practice, and Phase III Individual Assessment, to standardise panellists' interpretation of the CEFR levels, using exemplar tasks (Council of Europe 2005) already calibrated to the CEFR levels, and the CEFR scales during the Standardisation Training/Benchmarking procedure. After the judges are trained and reach a satisfactory level of agreement, they then allocate local test tasks and/or learners' performances to one of the CEFR levels and establish cut scores corresponding to the level.

Selection of well-qualified judges is the crucial first step to successful standard setting. Kaftandjieva (2004:28) suggests that qualified participants in standard setting are those who:

(a) are subject matter experts;
(b) have knowledge of the range of individual differences in the examinee population and be able to conceptualize varying levels of proficiency;
(c) are able to estimate item difficulty;
(d) have knowledge of instruction to which examinees are exposed;
(e) appreciate the consequences of the standards;
(f) collectively represent all relevant stakeholders.

To ensure quality and reliability of the results from Standard Setting, the *Standards* (American Educational Research Association et al 1999:54) suggests that 'a sufficiently large and representative group of judges should be involved to provide reasonable assurance that results would not vary greatly if the process were replicated.' The Manual (Council of Europe 2009:38) considers the minimum number of the panellists to be 12 to 15.

Although strenuous efforts are required during the process, standard setting does not intend to find 'preexisting or "true" cut score that separates real, unique categories on a continuous underlying trait (such as "competence")' (Cizek and Bunch 2007:18). Different standard setting methods may generate different results. To date no single standard setting method is considered to be suited to all conditions, and depending on methods applied, results from standard setting may vary. Jaeger (1989) classifies standard setting methods into two categories: (1) 'examinee-centred' if judgements are primarily about the test takers, and (2) 'test-centred' if cut score decisions are based on test content or test items. In examinee-centred methods, such as the Contrasting Groups method (Berk 1976) and the Borderline Group method (Zieky and Livingston 1977), the judgment is based on 'real' candidates; a panel of teachers who know their students well classify each student into pre-defined groups. In test-centred methods, such as the Tucker-Angoff method (Angoff 1971, Council of Europe 2009:61–66) and the Item-descriptor Matching Method (Ferrara, Perie and Johnson 2002), the panellists estimate the perceived item characteristics and classify the items based on the ability of defined or 'imaginary' borderline candidates. Among test-centred methods, the Basket Method (Alderson 2005) is often preferred due to its ease of implementation. The panel members place each test item into a basket which corresponds to one of the levels; no empirical information on the difficulty of the items is needed to present to the panel members. But care must be taken when applying the Basket Method since results from earlier studies suggest that it 'tends to produce lower (more lenient) standards than other methods' (Council of Europe 2009:76).

The Validation procedure, the final stage of the linking process, involves demonstrating evidence on the quality of the examination, procedural validity of the Standardisation Training and Standard Setting, internal validity of the Standard Setting, and external validation. The quality checking process involves the review of content coverage, trial testing and item analysis in test construction. The procedural validity of the Standardisation Training and Standard Setting assesses whether appropriate procedures are followed during familiarisation, standardisation training, and standard setting, in terms of explicitness, practicability, implementation, feedback, and documentation (Council of Europe 2009:95). The internal validity of the Standard Setting is concerned with the inter- and intra-judge consistency and the accuracy of the results, and external validation provides evidence from independent or external criteria to justify the results obtained from the linking procedures.

CEFR linking studies

Following the release of the preliminary version of the Manual (Council of Europe 2003), the Council of Europe called for participation from institutions and individuals in piloting the Manual. As a result, various institutions piloted the Manual and shared their experience on their linking activities. Some studies (Barni, Scaglioso and Machetti 2010, Khalifa et al 2010, Papageorgiou 2010, Wu and Wu 2010) applied one or two sets of procedures proposed in the Manual; others (Kecker and Eckes 2010, O'Sullivan 2008) undertook a systematic piloting of the Familiarisation, Specification, Standardisation, and Validation procedures suggested in the Manual.

Khalifa et al (2010) piloted the Familiarisation and Specification procedures in the Manual to relate four sections, i.e. the Listening, Reading, Writing, and Speaking papers, of First Certificate in English (FCE) to the CEFR B2 level, and provided reflections on their experience. Since FCE 'already has an established connection with the CEFR', the major aim of the study was to explore the possibility of incorporating the Manual procedures in FCE test development and validation processes to maintain the linkage relationships between Cambridge English exams and CEFR (Khalifa et al 2010:98). A total of 14 panellists participated in the study. The results showed that FCE was situated at the B2 level across all four skills. During the process, they supplemented 'non-Manual' activities, such as background reading on the relationships between Cambridge English exams and the CEFR, presentations on the CEFR, and feedback questionnaires. Khalifa et al (2010) reported finding some CEFR descriptors vague and difficult to relate to real-life experience and the Specification forms overlapping in recording information and, therefore, lacking conciseness and practicality.

Wu and Wu's (2010) study followed the 'internal validation' procedure, including Familiarisation, Specification, and Standardisation, to relate four levels of the reading components of the GEPT to the CEFR levels. A total of 15 panellists participated in the study. In the Specification phase, relevant forms in Chapter 4 of the Manual were applied to examine administrative procedures and text-level specifications of the GEPT reading tests. During the process, they incorporated the Dutch CEFR Construct Grid (Alderson et al 2006) to assess item-level comprehension and cognitive processing operations of the four levels of the GEPT. In addition to the qualitative analysis, sentence length, readability scores produced by the Dale-Chall (Chall and Dale 1995) and Fry (Fry 1968) formulas, and reading speed were calculated to quantitatively reflect and differentiate the four levels of the GEPT reading tests. The Standardisation Training and Standard Setting followed the three-phase procedure proposed in the Manual to train panellists to use the Basket Method (Alderson 2005) to relate their interpretations of the CEFR levels to the GEPT, using 12 CEFR scales relevant to reading, exemplary tasks calibrated to the CEFR levels (Council of Europe 2005), and locally produced GEPT reading tasks. The results showed that the GEPT Elementary, Intermediate, High-Intermediate, and Advanced levels were situated at CEFR A2, B1, B2, and C1 levels, respectively. Wu and Wu (2010) reported that relating the GEPT to the CEFR levels was difficult because the level descriptors do not explicitly define the quality of test takers' performance, the lexical and grammatical complexity of reading texts, and the test conditions that affect task difficulty, such as text length, and expected reading speed, at a particular CEFR level.

Kecker and Eckes's (2010) study and O'Sullivan's (2008) project both followed all four procedures, i.e. Familiarisation, Specification, Standardisation, and Validation, that the Manual proposed, to relate their exams to the CEFR. Kecker and Eckes's (2010) study examined the relationships of the four sections, i.e. listening, reading, writing, and speaking, of TestDaF and the CEFR at the B2 and C1 levels. The linking study began with Familiarisation and Specification procedures as the Manual suggested. To provide evidence on internal validity of TestDaF, the internal validation procedure described in the Manual, i.e. real-time verbal reports, task characteristics frameworks, and examiner and candidate feedback questionnaires, were adopted (Kecker and Eckes 2010:54). The external validity evidence was collected using Forms A9 to A22 in the Manual and through applying CEFR Grids for Speaking (Association of Language Testers in Europe 2005a) and Writing Tasks (Association of Language Testers in Europe 2005b) for analysis of test tasks. During the Standard Setting procedure, the Basket Method (Alderson 2005) was applied for the receptive skills and a modified variant of the benchmarking approach, focusing on individual assessments without discussion, for the productive skills to determine whether the candidate's ability was at the intended level for the tests. During the external validation, the German section of DIALANG and teacher judgment were used as external criterion measures for the receptive and the productive skills, respectively, to support the conclusions drawn from the previous procedures. Overall, Kecker and Eckes (2010) considered that the four-step methodology provided a pragmatic approach for the alignment purpose. Nevertheless, flaws were found in the CEFR scales, e.g. some parts of descriptors are inapplicable to the alignment since some examination content and the notion of task fulfilment are not covered in the CEFR scales. As regards the external criterion measures used in this study, the researchers reported that the DIALANG website was slow and unreliable from time to time, thus making it infeasible to use DIALANG to validate the linking relationships between the TestDaF and the CEFR levels. As to the exemplar tasks and performance samples that the Council of Europe provided, they were limited in number and format, and therefore, not suitable for the external validation purpose (Kecker and Eckes 2010:74).

O'Sullivan's (2008) project attempted to link City & Guilds B2 Communicator Level reading, writing, and listening components in English with the CEFR B2 level. This study stressed the importance of starting the linking process with a systematic and critical review of the quality of the exam in question to make sure that the exam is reliable and valid, and the significance of having experts from both within and outside of the exam board to undertake the review. In this study, the four-stage procedure proceeded in a linear manner, starting with Familiarisation, then Specification, Standardisation, and finally Validation. During the empirical validation procedure, learners took both the City & Guilds Communicator tasks and CEFR level exemplar tasks provided by the Council of Europe (2005). No writing performances were collected. The multi-faceted Rasch analysis of writing data collected and rated during the Standard Setting procedure was revisited. The results showed that the passing levels for the Communicator reading, listening and writing papers were in line with CEFR Level B2. Nevertheless, O'Sullivan (2008) considered the number of exemplar tasks and performances that the Council of Europe provided not only insufficient but also not representative enough to adequately reflect the range of proficiency at different CEFR levels. In addition, O'Sullivan (2008:85) suggested that 'limiting the validation evidence to estimates of internal and external validity is far too simplistic a view of validation.' He advocated collecting the validity evidence based on an explicit model of validation, such as Weir's (2005a) socio-cognitive validation framework, to provide theoretical justification behind the linking relationship.

Although the CEFR has been gaining widespread popularity and has had a positive impact on the practice of language testing since its publication, various case studies (Alderson (Ed) 2002, Figueras and Noijons 2009, Kecker and Eckes 2010, Khalifa et al 2010, Martyniuk and Noijons 2007, Morrow 2004, Wu and Wu 2010) discussed the difficulty in using the CEFR and called for further elaboration of level. For example, the CEFR recognises the importance of contextual features, but they are either not incorporated into the Can Do descriptors or lack explicit definitions, e.g. the range and frequency levels of grammatical structures and lexis, that differentiate the levels. The CEFR provides no guidance on what structures, lexis or other linguistic features learners might be expected to cope with in test tasks at various proficiency levels. Weir (2005b:12) commented that 'the CEFR provides little assistance in identifying the breadth and depth of productive or receptive lexis that might be needed to operate at the various levels.' Alderson et al (2006:13) shared the same view, noting that many of the terms in the CEFR are not explicitly defined (e.g. 'straightforward' and 'complex') and are comparative in nature (e.g. 'long' and 'longer'). Weir (2005b:2) argued that 'in its present form the CEFR is not sufficiently comprehensive, coherent or transparent for uncritical use in language testing.' Furthermore, Alderson, Figueras, Kuijper, Nold, Takala and Tardieu (2004:1) considered that the CEFR lacks 'sufficient theoretical and practical guidance to enable test specifications to be drawn up for each level.' Meanwhile, little advice is available on measures of quality check of the linking processes during the Validation stage. O'Sullivan and Weir (2011:18) suggested that 'the Manual fails to acknowledge advances in theoretical or practical validation from Messick (1980) to Weir [2005a]' and thus, a validation framework is needed to provide a theoretical basis for the CEFR linking process so that validity evidence can be generated in a more coherent fashion. Therefore, considerable additional resources are in need to establish claimed distinctions across different levels of the CEFR.

In light of the inherent weaknesses of the CEFR, other existing language proficiency scales will be reviewed next to identify criteria for better defining proficiency levels and help fill out the missing features in the CEFR.

Language proficiency scales

Numerous language proficiency frameworks or scales have been devised in different contexts to reflect 'a hierarchical sequence of performance ranges' (Galloway 1987:27). They divide language proficiency into defined levels which outline different degrees of achievement in terms that are meaningful to different users of the proficiency scales (Brindley 1986, 1991, Richterich and Schneider 1992, Trim 1977). Proficiency frameworks or scales which have gained wide recognition and have continued to be actively used include the ISLPR, formerly known as the ASLPR Scale (Ingram and Wylie 1979), the ACTFL Guidelines (Hiple 1987), the CLB (Pawlikowska-Smith 2000), and the CEFR (Council of Europe 2001). To explore features that are useful to define levels of language proficiency, a brief overview of the proficiency levels of the ISLPR, ACTFL, and CLB is presented as follows; the CEFR is detailed in the previous section.

The first version of the ISLPR, originally named the ASLPR, was released in 1979 to fulfil the Australian government's need for English as a Second Language (ESL) curriculum and materials for on-arrival adult immigrants from diverse language backgrounds. To reflect its increasing international popularity, the ASLPR was renamed the ISLPR in 1997. Its descriptors of reading, writing, listening, and speaking proficiency originated from the Foreign Service Institute (FSI) Scale (Foreign Service Institute School of Language Studies 1968), the most widely accepted scale available at that time. The ISLPR divides language proficiency into nine levels, i.e. Zero Proficiency (0), Initial Proficiency (0+), Elementary Proficiency (1-), Minimum Survival Proficiency (1), Survival Proficiency (1+), Minimum Social Proficiency (2), Minimum Vocational Proficiency (3), Vocational Proficiency (4), and Native-like Proficiency (5); three further levels, i.e. 2+, 3+, and 4+, are available but not explicitly defined. Descriptions of each level of the ISLPR focus on the language tasks that candidates can carry out and with what language forms these are carried out. Ingram (1990:47) indicated that 'to show gradation, some descriptive features are unavoidably comparative in nature and omission of a feature at one level that is included at the next implies that it is non-existent or insignificant at the lower level.' The ISLPR is now 'used in many different contexts ranging from education and the interpretation of test results to specifying migration regulations, in law courts, in classifying library material and in specifying the language skills required for vocational registration for teaching, nursing, and other vocations' (Ingram 2007:21).

The ACTFL Provisional Proficiency Guidelines were first published in 1982 under the Common Yardstick project 'to establish and implement second language proficiency guidelines for testing and for organising the language teaching curriculum' (Lantolf and Frawley 1985:337). The project also adopted the FSI oral proficiency testing procedure to describe proficiency of foreign language students and teachers in the USA. The ACTFL levels range from Novice Low (very basic proficiency) to Superior (native-like proficiency); the lower end of FSI scale, including levels 0 to 1, is subdivided into four levels, i.e. 0, Novice Low, Novice Mid, and Novice High, in the ACTFL Guidelines; the upper end of the FSI scale, including levels 3 to 5, is integrated into the Superior level. The sublevels of competency were defined according to the experience of language instructors and researchers. The Guidelines profile a hierarchy of integrated performance in speaking, listening, reading, and writing. Each ACTFL level consists of five components: function, content, context, accuracy, and text type. The ACTFL Guidelines now serve as the basis for the curriculum framework for foreign language instruction in the USA.

The Canadian Language Benchmarks (CLB) was initiated by Canadian government in 1992 to support the language learning needs of adult immigrants to the country. The CLB was developed based on Bachman's (1990), Bachman and Palmer's (1996), and Celce-Murcia, Dörnyei and Thurrell's (1995) models (Pawlikowska-Smith 2002:7). The CLB incorporates five components of communicative proficiency: linguistic competence, textual competence, functional competence, sociocultural competence, and strategic competence (Pawlikowska-Smith 2002:6). Communicative proficiency is divided into twelve benchmarks, BM1 to BM12, and arranged in three phases: Stage I Basic Proficiency, Stage II Intermediate Proficiency, and Stage III Advanced Proficiency. The CLB now serves to guide the teaching and assessment of ESL learners in Canada.

The ISLPR and the ACTFL Guidelines divide language proficiency into levels from 'Zero' to 'Native-like', while the CLB and the CEFR do not use an idealised native speaker as the norm (North 2000). The recurring features that the ISLPR, the ACTFL Guidelines, the CLB, and the CEFR use to differentiate levels of proficiency include vocabulary range, grammatical range, domain, degree of comprehension, content complexity, rhetorical organisation, genre, text length, ways of approaching reading texts, reading strategies, speed of reading, text abstractness, topic familiarity, cultural specificity, and subject specificity. See Table 1 for an overview of features that the ISLPR, the ACTFL Guidelines, the CLB and the CEFR use to establish levels.

Various features, such as linguistic (i.e. lexical and syntactic) and content complexity, and text length, are found repeatedly, if not consistently, in these scales. However, it appears that these proficiency scales define levels mostly through general statements on contextual features. These are certainly useful characteristics to describe levels, but alone they are not sufficient for the purpose of defining proficiency levels. McNamara (1996) and Weir (1993) suggested that the cognitive processes engaged by learners need to be given the same importance as contextual features so that test tasks and performance conditions can approximate to language use in the real world as closely as possible. To more comprehensively establish test levels, the following sections review studies on contextual impacts on reading performance and the cognitive processes underlying reading comprehension.

Table 1 Overview of features that the ISLPR, the ACTFL Guidelines, the CLB, and the CEFR use to establish levels

	ISLPR (1997)	ACTFL (1999)	CLB (2000)	CEFR (2001)
Purpose(s)	For developing new on-arrival programme in English as a second language for the Australian Adult Migrant Education Programme	To establish and implement second language proficiency guidelines for testing and for organising the language teaching curriculum	To describe the proficiency of adult immigrants in Canada and to define language ability globally in terms of features of 'real-life' performance	To provide a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc., across Europe
Proficiency levels	0 (Zero Proficiency) 0+ (Initial Proficiency) 1- (Elementary Proficiency) 1 (Minimum Survival Proficiency) 1+ (Survival Proficiency) 2 (Minimum Social Proficiency) 2+ 3 (Minimum Vocational Proficiency) 3+ 4 (Vocational Proficiency) 4+ 5 (Native-like Proficiency)	Novice Low Novice Mid Novice High Intermediate Low Intermediate Mid Intermediate High Advanced Low Advanced Mid Advanced High Superior	Stage I Basic Proficiency: BM1-BM4 Stage II Intermediate Proficiency: BM5-BM8 Stage III Advanced Proficiency: BM9-BM12	Basic User: A1 and A2 Intermediate User: B1 and B2 Proficiency User: C1 and C2

Table 1 (continued)

	ISLPR (1997)	ACTFL (1999)	CLB (2000)	CEFR (2001)
Vocabulary range	✓	✓	✓	✓
Grammatical range	✓	✓	✓	✓
Domain	Personal Social Academic Vocational	Personal Social Academic Professional	Community access Study/academic Workplace	Personal Public Educational Occupational
Degree of comprehension	✓	✓	✓	✓
Authenticity	✓	✓	–	–
Content complexity	✓	✓	✓	✓
Rhetorical organisation	✓	✓	✓	✓
Genre	✓	✓	✓	✓
Text length	✓	✓	✓	✓
Reading strategies	✓	✓	–	✓
Speed of reading	✓	✓	–	✓
Abstractness	✓	✓	√	✓
Topic familiarity	✓	✓	–	–
Cultural specificity	✓	✓	✓	–
Subject specificity	✓	✓	✓	✓
Handwriting	✓	–	–	✓

Contextual impacts on reading performance

To justify the use of language tests, we need to be able to generalise learners' test scores beyond their performance on the test to language use in the target language use (TLU) domain (Bachman and Palmer 1996:23–24). In order to establish an adequate correspondence between the test scores and their ability to use language in TLU situations, the characteristics of a given language test task have to reflect the features of a TLU task as much as possible (Bachman and Palmer 1996:23). Weir (1993:28–29) suggested that '[i]f the test tasks reflect real-life tasks in terms of important contextually appropriate conditions and operations it is easier to state what a student can do through the medium of English.' Various researchers (Alderson et al 2006, Bachman et al 1995, Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt and Schedl 2000, Fortus, Coriat and Fund 1998, Freedle and Kostin 1993, Khalifa and Weir 2009) have attempted to identify contextual features that affect performance in reading comprehension. Broad consensus on the features that are likely to impact on reading performance are detailed in Table 2.

Syntax and lexis have traditionally been considered important factors affecting reading comprehension (e.g. Alderson 2000, Bachman 1990, Grabe 2000, Khalifa and Weir 2009, Nuttall 1996, Perera 1984, Urquhart 1984, Urquhart and Weir 1998, Weir 1993). Read (2000) considered that lexical complexity is a strong predictor of text difficulty. On the other hand, Berman (1984) suggested knowledge of syntactic structures, such as parsing sentences into correct syntactic structures or identifying the constituent structures in sentences with complex syntax, such as words before the main verb, adverbial phrases before the main clause, and embeddings, is important in text understanding. Some research (e.g. Freebody and Anderson 1983, Haynes and Carr 1990, Nuttall 1996, Stanovich 2000, Urquhart 1984, Weir 1993) showed that vocabulary is more important for predicting reading test performance, while others (e.g. Alderson 1993, Bernhardt 1999, Shiotsu and Weir 2007) found that syntax correlates more strongly with reading performance. Still others (e.g. Barnett 1989) argued that both vocabulary and grammar affect reading comprehension to the same extent.

Many attempts have been made to develop procedures to estimate lexical and syntactical complexity. One common way to measure lexical difficulty is to check word frequency, i.e. how many words in the target text appear in a word frequency list, such as the British National Corpus (BNC) wordlist (BNC Consortium 2001) or Academic Wordlist (1998 and 2000). Since high-frequency words are generally identified faster than low-frequency words, texts containing more words from the high-frequency lists tend to be easier for readers to comprehend. Another simple estimator is word length, the number of letters or syllables a word contains, as shorter words tend to be more accessible, and therefore, text containing more short words is likely to be easier. The type-token ratio (TTR), the number of different words in a text,

Table 2 Contextual features that affect difficulty of reading comprehension

	Alderson et al (2006:25–30)	**Bachman et al (1995:49–50)**	**Enright et al (2000:14–38)**	**Fortus et al (1998:68–73)**	**Khalifa and Weir (2009:82–142)**
Task setting	–	–	–	–	–
General purpose	–	Purpose of the test	–	–	Overall test purpose
Target population	–	Participants	Participant	–	Test taker profiles
Structure of the test	–	Number of parts	–	–	(discussed under the response format)
Test focus	–	Nature of ability measured	–	–	Test focus
Communicative topic	Text source, topic	Topical characteristics	Content	Topic of text	Content knowledge
Authenticity	Authenticity (i.e. not abridged or simplified texts)	Naturalness	–	–	–
Time constraints	–	Time allotment/ degree of speediness	Amount of time allowed	–	Time constraints
Overall number of words	–		–	–	Overall number of words
Number of texts	–	Number of tasks	–	–	Number of texts
Expected speed of reading	–	Time of task	–	–	–

Table 2 (continued)

Scoring method	–	–	–	–	–
Knowledge of criteria	–	Explicitness of scoring criteria	–	–	Knowledge of criteria
Weighting	–	Relative importance of parts/tasks	–	–	Weighting
Text dimension	–	–	–	–	–
Domain	Domain	Target language use domain	–	Test domain	–
Genre	–	Register	Register		Discourse mode
Rhetorical structures/ discourse types	Discourse type	Rhetorical organisation	Rhetorical features	Rhetorical structure	Functional resources
Subject specificity	–	–	–	–	–
Cultural specificity	–	Cultural references	–	–	–
Abstractness	Nature of content (the abstraction dimension)	–	–	Level of abstractness	Nature of information
Lexical complexity	Vocabulary	Vocabulary	Vocabulary	Level of text vocabulary	Lexical variation/ frequency analysis
Syntactic complexity	Grammar	Syntax	Syntactic complexity	Level of grammatical complexity of text	Grammatical resources
Text length	Text length	Length of input	Amount of text	Length of text	Text length
Sentence length	–	–	–	–	Sentence length
Cohesion	–	Cohesion	Cohesion	–	–

Table 2 (continued)

	Alderson et al (2006:25–30)	**Bachman et al (1995:49–50)**	**Enright et al (2000:14–38)**	**Fortus et al (1998:68–73)**	**Khalifa and Weir (2009:82–142)**
Transition markers	–	–	Transition markers	–	–
Referentials	–	–	Antecedent reference	Number of referential markers	–
Lexical density	–	–	–	–	Lexical density (content words/total number of words)
Item dimension					
Response format	Response type	Type of expected response	Type of response formats	Multiple-choice questions	Response method
Content dimension	Recognise/evaluate; main idea/detail	–	Communicative purposes	Main idea/inference	–
Explicit dimension	Explicit/implicit	–	–	–	–
Globality	–	‘Narrow scope’: response made on the basis of a relatively limited amount of input	–	Very local to very global	–
Amount of processing	–	–	–	Little processing to great deal of processing	–

is also considered to be a useful index of text difficulty (Malvern and Richards 1997). A high TTR indicates a high degree of lexical variation which may take readers more time to process, and thus, usually suggests greater text difficulty. However, as the text gets longer, the number of word types falls. Therefore, when texts of different length are compared, standardised TTRs (STTRs), which calculate TTRs based on a fixed length of texts, are often applied instead of standard TTRs. As to syntactic complexity, sentence length is a convenient indicator. Berman (1984:153) suggested that 'efficient FL [foreign language] readers must rely in part on syntax to get at text meaning.' Generally speaking, texts with less complex grammar tend to be easier than those with more complex grammar, and short sentences are likely to contain simpler grammatical structures than long sentences. Text length is yet another potentially useful gauge of text difficulty. The longer the text readers have to process, the greater the language and content knowledge required, making reading more difficult. Grabe (2009:40) suggested that 'building up a general understanding of a longer text required more processing information than immediate word recognition, sentence parsing and propositional encoding.'

Readability indices are also commonly used measures of text difficulty. To date, more than 40 different readability formulas have been developed. The Flesch Reading Ease Score (FRES) and the Flesch–Kincaid (FK) Grade Level index are the most popular ones among researchers in the field of education. Both the two formulas use the same measures, i.e. word length and sentence length, but the two variables are assigned different weightings. The FRES ranges from 0 to 100; a higher score indicates that the text is easier to read while a lower number suggests that the text is more difficult to read. The FK index converts the FRES to a US grade level, and the readability level of texts can be interpreted straightforwardly based on the number of years of education for learners to receive in the US in order to understand the text. Coh-Metrix L2 Reading Index (Crossley, Greenfield and McNamara 2008), in addition to lexical and syntactic features of reading texts, takes textual coherence into account to assess text comprehension. Coh-Metrix readability scores are reported on a scale of 0 to 30; a higher score indicates easier readability.

In addition to lexical and syntactic complexity, Alderson et al (2004:127) suggested 'abstract information often implies a linguistic complexity that may further stretch the L2 reader's resources.' In general, abstract texts are harder to understand than texts describing real objects, events, or activities (Alderson 2005), and abstract words are more difficult to process than concrete words (Anderson 1974, Corkill, Glover and Bruning 1988).

The effect of text cohesion on comprehension is not as straightforward as that of lexical and syntactic complexity. Alderson (2000:68) noted that cohesion effects on comprehension are relatively weak, probably because the effects of text topic and reader's language proficiency mediate with the impact of cohesion, and therefore, lack of connectives might not influence comprehension to a great extent. On the other hand, Goldman and Rakestraw (2000) found explicit cohesive devices to contribute positively to establishing textual coherence, and coherent texts tend to be easier to comprehend than less coherent texts (Beck, McKeown, Sinatra and Loxterman 1991). McKeown, Feiner, Robin, Seligmann and Tanenblatt (1992) also suggested that text coherence contributes substantially to comprehension when the content is relatively unfamiliar to the readers; in the meantime, coherent texts also enable readers with relevant background knowledge to understand texts better.

Unlike linguistic complexity, text length, text organisation, cohesion and coherence, the effect of text topics and text types or genre on text comprehension have not yet been thoroughly researched and are not clearly understood (Nuttall 1996:221). Although it is generally considered that the more knowledge of text topic readers have, the easier it is for them to process the text, Alderson (2000:69) argued that 'topic (un)familiarity cannot be compensated for by easy vocabulary: both difficult vocabulary and low familiarity reduce comprehension, but texts with difficult vocabulary do not become easier if more familiar topics are used, and vice versa.' Urquhart and Weir (1998:143) considered it important to cover content that test takers are sufficiently familiar with so that schemata to employ appropriate skills and strategies to comprehend the text can be activated. As both subject areas (Hughes 1989:93) and culturally specific content (Sasaki 2000) may affect reading comprehension, special attention should also be paid to the subject and cultural specificity of the texts used in a reading test.

Text types or rhetorical features refer to 'one of the traditional discourse models of narration, description, exposition, and argument/persuasion' (Weigle 2002:62). Although how text types create difficulty for readers is not yet well understood, Alderson (2000:39–40) argued that:

> Knowing how texts are organised – what sort of information to expect in what place – as well as knowing how information is signalled and how changes of content might be marked – has long been thought to be of importance in facilitating reading. For example, knowing where to look for the main idea in a paragraph, and being able to identify how subsidiary ideas are marked, ought in principle to help a reader process information.

Barnett (1989:56) considered it important to examine the impact of text type or structure on text difficulty. To avoid bias in test performance, Nuttall (1996:221) suggested including a variety of different text types in a reading comprehension test.

Genre is defined as 'the expected form and communicative function of the written product; for example, a letter, an essay, or a laboratory report' (Weigle 2002:62) and takes in 'salient features and conventions which are shaped by communicative purposes' (Hyland 2000:62). Genre also has potential impact on test performance; a particular genre involves specific conventional (lexical, syntactical, semantic, and discoursal) features which are likely to affect text processing (Bhatia 1997, Hyland 2000). Therefore, when selecting genres for a test, special attention should be paid to ensure that they are at an appropriate level of specificity and are not culturally biased or do not favour any group of the test population (Weir 1993).

In addition to the range of text variables that affect comprehension difficulty, the performance conditions, such as time constraints and response formats, also influence how learners process the reading texts. For example, Nuttall (1996:56) reported finding that ESL university students read at approximately 200 words per minute (wpm), but when they are studying texts that are difficult for them the speed might drop to as slow as 60 wpm, while university students whose native language is English read at a wide range of reading rates (300 to 800 wpm). Weir (2005a:65) suggested that 'timing clearly impact[s] on the processing and hence on the theory-based validity'; if more than enough time is allowed to complete an expeditious task, test takers tend to use careful reading instead of quick selective reading, and therefore test constructs may be distorted. The response format is also an important performance condition that affects results of reading comprehension (Alderson et al 2006, Bachman et al 1995, Enright et al 2000, Khalifa and Weir 2009). For example, multiple-choice questions may create very different comprehension and response processes (Embretson and Wetzel 1987) and they might activate different reading processes. Rupp, Ferne and Choi (2006:441) reported finding that 'learners view responding to multiple-choice questions as a problem-solving task rather than a comprehension task.' The impact of response format on level differentiation is not yet fully understood. Thus, Khalifa and Weir (2009:83) suggested that it will be useful to survey examination board practice to determine to what extent test formats help make distinctions between levels.

Based on the literature review of previous research on contextual impacts on reading performance in this chapter, contextual features that may be useful in describing level distinctions for this study are identified in Table 3. Empirical studies are needed to explore relative degrees of complexity and the range of the contextual features in terms that are specific enough to distinguish levels with sufficient precision (Bachman and Savignon 1986:388). The methods to be applied in this study are discussed in Chapter 3.

Cognitive demands are now discussed to determine relevant cognitive processing parameters that are considered useful in differentiating between test levels in reading tests.

Table 3 Contextual features selected to be analysed in this study

Task setting	Text dimension
General purpose	Text length
Target population	Text type
Structure of the test	Genre
Test focus	Rhetorical structures/discourse types
Communicative topic	Subject specificity
Time constraints	Cultural specificity
Overall number of words	Abstractness
Number of texts	Lexical complexity
Expected speed of reading	Word frequency
Item dimension	Word length
Response format	Type-token ration
Amount of processing	Readability (e.g., FR Ease, FK Grade Level, Coh-Metrix)
	Syntactic complexity
	Sentence length
	Text cohesion
	Connectives
	Referentials
	Lexical density

Cognitive processing in reading

Cognitive processes underlying language use and test performance have received considerable attention among researchers in the field of language testing since the 1970s. The interaction between cognitive processing and second language use has been incorporated in recent models and theories of second language proficiency, such as Bachman's (1990), Bachman and Palmer's (1996), Canale's (1983), Canale and Swain's (1980), and Weir's (2005a) models. Canale and Swain (1980) were the earliest researchers to introduce *strategic competence* to the field of language testing. Canale and Swain's framework of communicative competence included grammatical competence, i.e. the knowledge of grammar, lexis, morphology, syntax, semantics and phonology; sociolinguistic competence, i.e. the knowledge of the sociocultural rules of language use; and strategic competence, i.e. knowledge of communication strategies that can be employed to compensate for breakdowns in communication due to insufficient competence in one or more components of communicative competence (1980:29–30). Canale (1983:339) later refined Canale and Swain's framework by extending the definition of strategic competence further to 'enhance the rhetorical effect of utterances' and adding discourse competence as 'to combine and interpret meanings and forms to achieve unified text in different modes (e.g. casual conversation, argumentative essay, or recipe).' In both Canale and Swain's (1980) and Canale's (1983) models, the main function of strategic competence was to facilitate communication, but they mentioned little about the mechanisms by which strategic competence operates (Bachman 1990:99). Building on Canale and Swain's framework and Canale's model, Bachman's (1990) model of communicative language ability outlined the interrelationships between different competence components, with strategic competence playing a central role by mediating other components of communicative language ability, i.e. language competence and psychophysiological mechanisms. In his model, strategic competence, defined as 'the capacity that relates language competence, or knowledge of language, to the language user's knowledge structures and the features of the context in which communication takes place' (1990:107), performs three functions, i.e. assessment, planning, and execution, to achieve a communicative goal. Bachman and Palmer's (1996) model further highlighted the processing and contextual issues and indicated directions of the dynamic and interactive relationships between metacognitive strategies, i.e. goal setting, assessment, and planning, and language users' personal characteristics, topical knowledge, language knowledge, and affective schemata. The model facilitated systematic evaluation of constructs of test tasks; however, it did not present a coherent picture of how individual components come into play or how they affected language performance. Weir's (2005a) socio-cognitive validation framework advanced Bachman and Palmer's model by conceptualising the relationships among test taker characteristics, the contextual characteristics of TLU tasks, language knowledge, and cognitive processing skills; within each component, distinct elements were identified for researchers to collect evidence and examine various aspects of test validity.

Researchers (e.g. Davis 1968, Jang 2009, Khalifa and Weir 2009, Lumley 1993, Munby 1978, Weir and Porter 1996) have long attempted to identify reading skills or subskills under various performance conditions. Davis (1968) devised items to test the eight skills he identifies, and Munby (1978) compiled a list of 'microskills' that he considered to contribute to readers' abilities to understand texts. Results from earlier empirical studies (e.g. Einstein, McDaniel, Owen and Cote 1990, McDaniel, Blischak and Einstein 1995, McDaniel, Anderson, Einstein and O'Halloran 1989, McDaniel, Einstein, Dunay and Cobb 1986, Urquhart and Weir 1998:96) showed that readers employed different skills and strategies and thus different processing activities were involved when they read for different purposes across different types of texts. For example, when they read newspapers or advertisements, they tend to skip passages and ignore details that are not relevant to their interest or needs. On the other hand, when they read a single text or multiple texts for learning purposes, or when they read for general purposes, different cognitive operations may be elicited (Goldman 1997, Perfetti, Rouet and Britt 1999).

Based on earlier research, Weir and Porter (1996) and Urquhart and Weir (1998) classified reading processes and skills into four broad categories:

(a) Expeditious reading at the global level, i.e. skimming for the gist and searching for information.
(b) Expeditious reading at the local level, i.e. scanning for specific information through word-matching strategies.
(c) Careful reading at the global level, i.e. understanding explicitly stated main ideas, inferring propositional meanings and pragmatic meanings.
(d) Careful reading at the local level, i.e. inferring lexical meanings and understanding syntax.

Khalifa and Weir (2009) further decomposed cognitive processes in reading into eight hierarchical levels: word recognition, lexical access, syntactic parsing, establishing propositional meaning, inferencing, building a mental model, creating a text level representation, and creating an intertextual representation. Khalifa and Weir's model of reading attempted to relate underlying abilities, i.e. text structure knowledge (genre and rhetorical tasks), general knowledge of the world, topic knowledge, syntactic knowledge, and lexical knowledge, to performance and processing conditions and presented cognitive processes in a sequential frame to profile the impact of language knowledge on different levels of processing.

When defining the constructs of examinations at different proficiency levels, it is useful to obtain evidence on how examinees achieve various types of reading comprehension from cognitive processing perspectives. Weir (2005a:18) suggested that for a test task to be valid, the language processing which underlies the operations in test conditions should replicate that required in real-life language use as far as possible. The CEFR overlooks the role of cognitive operations in defining the different levels of the framework (Alderson 2007:661). Khalifa and Weir (2009:82) suggested that it would be helpful to survey potential formats for testing reading at different levels and investigate cognitive processing operations they are likely to activate.

Test comparability

Test users, for various purposes, often express a need for information about how scores from different tests relate to one another. One approach to score comparisons discussed previously is to examine how test takers' performance on two different tests relate to each other, and to what extent their scores are correlated to each other. Making comparisons this way tends to focus solely on the notion of score equivalence and this is generally regarded as insufficient in the language testing context, since ‘each test is designed for a different purpose and a different population, and may view and assess language traits in different ways as well as describing test-taker performance differently’ (Davies, Brown, Elder, Hill, Lumley and McNamara 1999:199). Thus, relevant factors affecting test scores need to be taken into account when score comparisons between different tests or levels are made (Geranpayeh 1994:62). Taylor (2004:3) suggested that in addition to score comparison, careful thought must be given to various features of the tests, such as:

> . . . purpose, construct definition, test method, content breadth and depth, skills coverage, accuracy of measurement, predictive/diagnostic power, score interpretability, test length, accessibility, . . . cost, . . . degree of specificity, currency and recognition, relationship to curriculum, [and] impact in the wider world.

A recent alternative to score comparison is to place scores from different tests on a common scale or within a common framework of reference which summarises features that appear repeatedly across tests. Thus, a convenient point of reference or readily interpretable results can be provided to meet test users’ demand for score comparability.

Using language frameworks as a medium for test comparison may seem appealing due to its ease of interpretation for stakeholders. Nevertheless, Taylor (2004:4) argued:

> . . . comparative frameworks cannot easily accommodate the multidimensional complexity of a thorough comparative analysis; the framework will focus on shared elements but may have to ignore significant differentiating features . . . The result is likely to be an oversimplication and may even encourage misinterpretation on the part of users about the relative merits or value of different exams . . . . [T]here is always a danger that they are adopted as prescriptive rather than informative tools.

Strenuous efforts need to be made to align the exams to a defined level of the framework and validate the distinctions between tests at different proficiency levels. There should be explicit procedures, both qualitative and quantitative, incorporating criterial features that affect comprehension performance for test validation and test comparison.

Building on the findings from the literature review, the following chapter addresses research methods that can be used to investigate such a comparison in a comprehensive and principled manner. Chapter 3 presents an empirical framework for test validation and test comparability in terms of contextual features, cognitive operations, and test results.

3 Methodology

Chapter overview

This chapter describes and explains the methodology for validating an L2 level-based reading test system, the GEPT, developed for use in Taiwanese educational contexts. The process involved both internal validation of the GEPT to demonstrate differentiation across different GEPT levels and external validation to establish equivalence between the GEPT and an alternative CEFR-aligned test, i.e. the core Cambridge English examinations, at the same CEFR level. *Vertical* comparisons of the GEPT at CEFR B1 and B2 levels, in terms of test results, contextual and cognitive processing parameters, were made to generate data to answer Research Question 1: Is a GEPT reading test designed to measure at CEFR B2 level more difficult than a GEPT reading test designed to measure at CEFR B1 level in terms of test results, contextual parameters, and cognitive processing skills? *Horizontal* comparisons between the GEPT and Cambridge English reading tests at the same CEFR level, in terms of test results, contextual and cognitive processing parameters, were drawn to produce data to answer Research Question 2: Are GEPT reading tests at CEFR B1 and B2 levels comparable to alternative CEFR-linked measures in terms of test results, contextual parameters, and cognitive processing skills? Empirical procedures to establish vertical differentiation across exams targeting different proficiency levels within the GEPT test battery are first discussed. The GEPT was linked vertically in terms of test scores to examine whether a GEPT targeting a higher level, i.e. CEFR B2 level, was more difficult than a GEPT reading test targeting a lower level, i.e. CEFR B1 level. Statistical procedures for comparison between tests which are developed by different exam boards targeting the same proficiency level are then outlined. The GEPT reading tests at the B1 and B2 levels were compared horizontally with core Cambridge English examinations to assess whether the two different reading tests at the same CEFR level were comparable. Cambridge English Reading papers at the B1 and B2 levels were selected as external criterion measures since earlier research (e.g. Kecker and Eckes 2010, O'Sullivan 2008) suggested other measures, such as DIALANG and exemplar tasks provided by Council of Europe (2005), were not adequate for the intended purpose. The core Cambridge English examinations are one of the few examination suites that have made claims about the relationship of the examinations to the levels of the CEFR, and were therefore adopted in this study to fill this void.

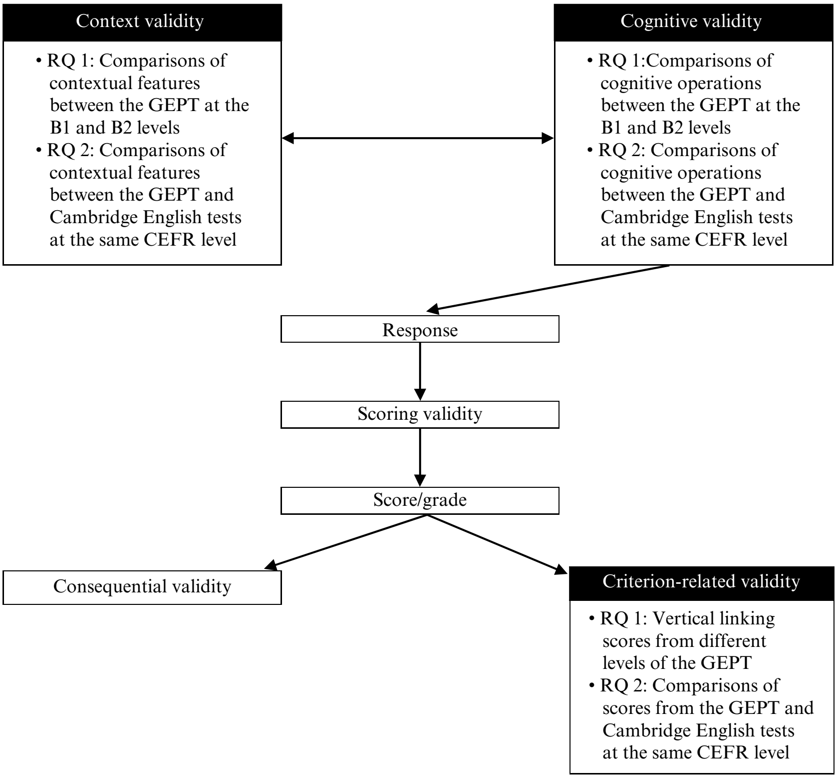
Following the statistical comparison of test results, the contextual and cognitive parameters of the GEPT tests at the B1 and B2 levels were examined and compared with those in the two Cambridge English tests at equivalent levels. The contextual features, identified in Chapter 2, that may affect comprehensibility and difficulty of reading tasks were used to develop methodological approaches for collecting evidence on context validity in the two sets of exams. The cognitive operations identified in Chapter 2 were used to design instruments which could be applied to both sets of tests to gather evidence on cognitive validity from both experts' and test takers' perspectives. Based on the results from the analysis of contextual features and cognitive operations, parameters that are useful to explicitly differentiate difficulty levels were then identified.

The methodology for generating data on the context, cognitive, and the criterion-related validity of the GEPT and comparable CEFR-aligned tests from Cambridge English drew on Weir's (2005a) socio-cognitive validation framework. To answer Research Question 1 (Is a GEPT reading test designed to measure at CEFR B2 level more difficult than a GEPT reading test designed to measure at CEFR B1 level in terms of test results, contextual parameters, and cognitive processing skills?) scores from different levels of the GEPT were first vertically linked and placed on a common score scale to provide statistical evidence on criterion-related validity, and further data were generated through comparing contextual features and cognitive operations between the GEPT at the B1 and B2 levels, respectively, to provide context validity and cognitive validity evidence to demonstrate the existence of differences in difficulty between the two GEPT levels.

To answer Research Question 2 (Are GEPT reading tests at CEFR B1 and B2 levels comparable to alternative CEFR-linked measures in terms of test results, contextual parameters, and cognitive processing skills?), horizontal comparisons between contextual features, cognitive operations, and test results of the GEPT and Cambridge English reading tests at the same CEFR level were made to provide both qualitative and quantitative evidence on context validity, cognitive validity, and criterion-related validity, respectively. Figure 1 below visually summarises how the current research design relates to Weir's (2005a) socio-cognitive framework.

First, the methodology for the statistical vertical scaling of GEPT tests is detailed below.

Figure 1 Validation procedures for GEPT level differentiation based on Weir's (2005a) socio-cognitive framework (RQ= Research Question)



Context validity
• RQ 1: Comparisons of contextual features between the GEPT at the B1 and B2 levels
• RQ 2: Comparisons of contextual features between the GEPT and Cambridge English tests at the same CEFR level
Cognitive validity
• RQ 1:Comparisons of cognitive operations between the GEPT at the B1 and B2 levels
• RQ 2: Comparisons of cognitive operations between the GEPT and Cambridge English tests at the same CEFR level
Response
Scoring validity
Score/grade
Consequential validity
Criterion-related validity
• RQ 1: Vertical linking scores from different levels of the GEPT
• RQ 2: Comparisons of scores from the GEPT and Cambridge English tests at the same CEFR level

Vertical scaling of scores from GEPT reading tests at different levels

To answer Research Question 1, the GEPT level framework was first examined through vertically scaling scores from GEPT reading tests at different levels to establish whether the projected increases in difficulty were reflected in terms of CEFR levels and test results. The GEPT level framework was examined internally in terms of difficulty. Scores from tests at different GEPT levels were linked and placed on a common scale to determine empirically how far apart each GEPT level was from its adjacent levels, in terms of common scale score units.

Vertical scaling is a process to place scores obtained from tests of different difficulty onto a common score scale and provides a systematic evaluation of level differentiation within a multilevel test battery. It provides evidence that a GEPT test at the B2 level is more difficult than a GEPT test at the B1 level. In this study, IRT Rasch model estimation was used to scale scores from reading tests at different GEPT levels, i.e. Elementary, Intermediate, and High-Intermediate levels, onto a 'vertical scale' (Tong and Kolen 2007:228) to empirically validate the projected increase in difficulty across test levels. The results of the IRT analysis determined how far apart each GEPT level was from its adjacent levels, in terms of common scale score units. See Table 4 for an overview of the GEPT Elementary, Intermediate, and High-Intermediate reading tests.

Table 4 Overview of the GEPT Elementary, Intermediate, and High-Intermediate reading tests

GEPT level	Part	Task type	No. of items		Time (minutes)
Elementary	1	Vocabulary and structure	15	35	35
	2	Cloze	10		
	3	Reading comprehension	10		
Intermediate	1	Vocabulary and structure	15	40	45
	2	Cloze	10		
	3	Reading comprehension	15		
High-Intermediate	1	Vocabulary and structure	10*	45	50
	2	Cloze	15		
	3	Reading comprehension	20		

**The number of questions in Part 1 has been reduced from 15 to 10 since 2010.*

The selection of samples in this study followed the general requirements noted by Patz and Yao (2007) for test forms and examinee populations used to produce vertical scales. To make the linking relationship more robust, the reading testlets contained over 50% of the total number of items as vertical anchors, i.e. items in common. This proportion was greater than what has been suggested as adequate for linking purposes; e.g. Patz (2007:12) suggested using at least 15 items or more and Hanson and Beguin (2002:5) used 20 items out of 60 items.

Participants

A total of 827 target test takers at the GEPT Elementary, Intermediate, and High-Intermediate levels participated in the linking study. Target test takers of the Elementary level, i.e. CEFR A2 level, are those whose English proficiency is equivalent to that of junior high school graduates in Taiwan, aged 14 to 16; those of the Intermediate level, i.e. CEFR B1 level, are equivalent to senior high school graduates in Taiwan, aged 17 to 19; and those of the High-Intermediate level, i.e. CEFR B2 level, are equivalent to university graduates of non-English majors in Taiwan, aged 19 to 23.

Instruments

This study employed common-item non-equivalent groups design. Tong and Kolen (2007) suggested when common item design is applied, using the middle level, instead of the lowest or highest level, as internal anchor, may reduce the extent of scale shrinkage. Test questions from the GEPT Elementary, Intermediate, and High-Intermediate level reading tests were selected to form shortened versions of the GEPT tests of each level. The statistical characteristics (i.e. mean and spread of the item difficulties) of each shortened version at the specified level were roughly the same as those of the operational tests of the same level. The Elementary test set contained 15 Elementary reading test items, the Intermediate test set 25 Intermediate reading test items, and the High-Intermediate test set 19 High-Intermediate reading test items. The three test sets were then grouped into two testlets: Testlet 1 was composed of the Elementary and the Intermediate test sets, and Testlet 2, the Intermediate and the High-Intermediate test sets. The Intermediate test set (see the shaded cells in Table 5) was embedded in both testlets as an internal anchor, functioning as a basis for linkage. The anchor items were ordered in the same way in each testlet in which the items appeared. Items in both the Elementary and High-Intermediate test sets were located on the scale of the Intermediate item parameter estimates.

Table 5 Number of items per reading testlet for vertical linking

Testlet	No. of Items	Elementary	Intermediate	High-Intermediate
1	40	15	25	0
2	45	0	25+1*	19

**The extra one Intermediate test item in Testlet 2 is to make the total number of test items the same as the operational GEPT High-Intermediate reading test in order to approximate the test condition.*

To approximate the test conditions of the operational GEPT, the total number of items in each testlet was set as close to that in the operational GEPT reading tests as possible. The two testlets were then administered to two groups of students at different levels of English proficiency. Linkage was established through test takers' performance on the anchor items.

Procedures for data collection

Target examinees of Elementary and Intermediate levels, 429 students aged 14 to 17, were invited from three high schools to take Testlet 1, and target examinees of Intermediate and High-Intermediate levels, 398 students aged 17 to 20, were invited from one high school and three colleges to take Testlet 2. See Table 6.

Table 6 Vertical scaling data collection design

Examinee group	Testlet by GEPT Level		
	Elementary	**Intermediate**	**High-Intermediate**
Group 1 (N=429)	Testlet 1		
Group 2 (N=398)		Testlet 2	

Data analyses

This study employed BILOG-MG (Zimowski, Muraki, Mislevy and Bock 2003), using the marginal maximum likelihood estimation (MMLE) mode with the group option, to scale the Rasch model item and ability parameter estimates across three levels of the GEPT reading tests. When the non-equivalent groups design is used for vertical scaling, BILOG-MG generally performs better on item and ability estimation than other IRT estimation programs, e.g. WINSTEPS, which are not based on multiple groups (Camilli 1988, Camilli et al 1993, Custer et al 2006, DeMars 2002). BILOG-MG uses MMLE and has a group option during estimation, while WINSTEPS uses joint maximum likelihood estimation (JMLE) and does not have a group option. Earlier research (e.g. Camilli et al 1993, Custer et al 2006, Skaggs and Lissitz 1985, Williams et al 1998) showed that ability scales created with MMLE were less prone to measurement error and less affected by the range restriction encountered in vertical scaling than maximum likelihood estimation (MLE) or JMLE.

Both concurrent estimation and separate estimation were performed and then compared, although few significant differences between parameters estimated with the two methods were observed in previous research (Camilli et al 1993, Hanson and Beguin 2002). Concurrent calibration involves estimation of item parameters for items in both testlets at the same time; thus, the estimates for the common (Intermediate level in this study) items are based on a larger sample size as responses from both test administrations are included. Separate calibration obtains item parameter estimates for one testlet (Testlet 1 in this study) and then estimates parameters in the other form (Testlet 2 in this study) with the common item parameters fixed at their estimated values using the first form. Previous studies showed that concurrent calibration was more accurate when the data fit the IRT model (Kim and Cohen 1998), but separate calibration was more robust to violations of the IRT assumptions due to multidimensionality (Kim 2007). In this study, concurrent estimates were used operationally, while separate estimates were computed to triangulate the results from concurrent estimation.

Next, the methodology for the horizontal statistical comparison of the GEPT and Cambridge English test scores is explained.

Comparisons between scores from GEPT and Cambridge English reading tests at the same CEFR level

An important component of test validation is the extent to which scores on a test are comparable with scores obtained on an established measure which aims at the same population. Such evidence serves as evaluation criteria and helps convince test users that the test under review is appropriate to the level it intends to measure at.

To determine whether the GEPT and Cambridge English reading tests targeting the same level were comparable in terms of test takers' performance, the GEPT reading tests at CEFR B1 and B2 levels were horizontally related to the Cambridge English tests at the same CEFR level. See Table 7 for an overview of the GEPT and Cambridge English reading tests at the B1 and B2 levels.

In this study, CTT statistics were used to compare GEPT target examinees' performances on the GEPT and Cambridge English reading tests at the same CEFR level to determine whether the two pairs of tests were equivalent in terms of test results. In addition to psychometric characteristics, non-psychometric properties of the GEPT and Cambridge English reading tests, i.e. tests' contextual features and test takers' cognitive operations, were also analysed; see the latter part of this chapter.

Participants

A total of 132 target examinees of the GEPT Intermediate level, targeting CEFR B1 level, took part in this study. Selection criteria included those who took the GEPT Elementary tests within a year and scored over 100 points (Elementary high-pass test takers; the passing score for the GEPT reading tests at all levels is set at 80 out of 120), and also those who took the GEPT Intermediate tests within a year and scored between 60 and 80 points (Intermediate near-pass examinees).

Another 138 target examinees of the GEPT High-Intermediate level,

Table 7 Overview of the GEPT and Cambridge English reading tests at CEFR B1 and B2 levels

CEFR level		Part	Item type	No. of Items		Time
B1	**GEPT Intermediate**	1	Fifteen incomplete sentences, each followed by four options	15	40	45 minutes
		2	Two texts; each contains five missing words or phrases; beneath each text, there are five items, each with four options	10		
		3	Five texts; each text followed by two to four comprehension questions with four options	15		
	Cambridge PET	1	Five short discrete texts, comprising signs and messages, postcards, notes, emails, labels, etc., each followed by a comprehension question with four options	5	35	50 minutes
		2	Five items in the form of descriptions of people to match to eight short adapted authentic texts	5		
		3	Ten statements about a longer factual/informational text and test takers decide whether they are true or false	10		
		4	A text containing attitude/opinion followed by five multiple-choice questions with four options	5		
		5	Ten items, with an adapted-authentic text, of a factual or narrative nature	10		
B2	**GEPT High-Intermediate**	1*	Ten incomplete sentences, each followed by four options	10*	45	50 minutes
		2	Two texts; each contains seven to eight missing words or phrases. Beneath each text, there are seven to eight items, each with four options	15		
		3	Four texts and one chart; each text or chart is followed by two to six comprehension questions with four options	20		
	Cambridge FCE	1	A text followed by eight multiple-choice questions.	8	30	60 minutes
		2	A text from which seven sentences have been removed and placed in a jumbled order, together with an additional sentence, after the text	7		
		3	A text or several short texts preceded by fifteen multiple-matching questions	15		

**The number of questions in Part 1 has been reduced from 15 to 10 since 2010.*

targeting CEFR B2 level, participated in this study – the selection criteria was the same as for the 132 examinees described previously.

Instruments

Reading papers of GEPT Intermediate level (Language Training and Testing Center 2011a) and Cambridge PET (Cambridge ESOL 2009) and those of GEPT High-Intermediate level (Language Training and Testing Center 2011b) and Cambridge FCE (Cambridge ESOL 2007) were used to investigate the relationships between the two exams at CEFR B1 and B2 levels, respectively, in terms of test takers' performance.

Procedures for data collection

A single group design was used. Test takers took both the GEPT and Cambridge English tests at the same level and were randomly divided into four groups: Groups 1 and 2 took reading tests only, and Groups 3 and 4 were assigned to take reading tests and fill out the Cognitive Processing Checklist (see Appendix 1 and Table 17 in this chapter) immediately after they answered each comprehension question. To minimise any practice effect, the order of administering the GEPT and Cambridge tests was counterbalanced: Group 1 took the GEPT first and then took the Cambridge English test, while Group 2 took the Cambridge English test first and then the GEPT. The order of administration for Groups 3 and 4 was also counterbalanced (see Table 8 for data collection design).

Data analysis

Test results from the GEPT and Cambridge English reading tests at the same CEFR level were compared. Means, standard deviations, overall percentages of items correct, score distributions, Pearson product moment correlation coefficients between the GEPT and Cambridge English tests at the same CEFR level, and t-tests of differences in means between the two pairs of the tests were computed to investigate the empirical relationships between the two CEFR-aligned reading tests. As well as statistical comparison between the tests, the construct validity parameters of the tests under review, i.e. their context and cognitive validity, need to be considered in more depth. We next turn to the contextual parameters of both sets of GEPT and Cambridge English tests at the B1 and B2 levels. Procedures to compare GEPT exams with equivalent Cambridge English examinations across and between levels to generate evidence on this validity component will be described.

Table 8 Horizontal comparison data collection design

	Reading tests only		Reading tests and the Cognitive Processing Checklist	
B1 level (N=132)	**Group 1 (N=21)**	**Group 2 (N=40)**	**Group 3 (N=36)**	**Group 4 (N=35)**
Test Session 1	GEPT Intermediate in 45 minutes	Cambridge English PET in 50 minutes	GEPT Intermediate and the checklist in 55 minutes	Cambridge English PET and the checklist in 60 minutes
Break	10 minutes	10 minutes	10 minutes	10 minutes
Test Session 2	Cambridge English PET in 50 minutes	GEPT Intermediate in 45 minutes	Cambridge English PET and the checklist in 60 minutes	GEPT Intermediate and the checklist in 55 minutes
B2 level (N=138)	**Group 1 (N=31)**	**Group 2 (N=34)**	**Group 3 (N=31)**	**Group 4 (N=42)**
Test Session 1	GEPT High-Intermediate in 50 minutes	Cambridge English FCE in 60 minutes	GEPT High-Intermediate and the checklist in 60 minutes	Cambridge English FCE and the checklist in 75 minutes
Break	10 minutes	10 minutes	10 minutes	10 minutes
Test Session 2	Cambridge English FCE in 60 minutes	GEPT High- Intermediate in 50 minutes	Cambridge English FCE and the checklist in 75 minutes	GEPT High-Intermediate and the checklist in 60 minutes

Contextual parameter analysis: Vertical and horizontal comparisons of the GEPT and Cambridge English exams at the B1 and B2 levels

The contextual parameters of a reading test will contribute to the difficulty of that test in terms of their effect on the cognitive load they place on processing. In GEPT tests, we would expect the contextual difficulty indices to be higher in a B2 level examination than in a B1 level examination. Similarly when we compare two tests deemed to be at a comparable level we would expect a good degree of similarity between the contextual difficulty indices in each. This section describes the methodology for establishing evidence in respect of these contextual parameters.

Textual features that affect the comprehensibility and difficulty of reading tasks, identified in Chapter 2, were analysed through both automated tools and expert judgement. Traditionally, contextual features are analysed based on experts' holistic interpretation. Advances in automated textual analysis have made it possible to examine analytically on a wider range of textual characteristics to complement human judgement.

Instruments for automated analysis of contextual features

In this study, reading texts from six GEPT and six Cambridge English reading papers were analysed. The GEPT texts at the B1 level were taken from Intermediate Level Past Papers 3, 4, and 5 (Language Training and Testing Center 2005, 2009a, 2011a), and texts at the B2 level taken from High-Intermediate Level Practice Paper (Language Training and Testing Center 2010) and Past Papers 4 and 5 (Language Training and Testing Center 2009b, 2011b). The Cambridge English test papers were those published in the public domain and intended to reflect the content and difficulty of the operational tests, including texts at the B1 level taken from three Reading papers in the PET Handbooks for teachers (Cambridge ESOL 2004, 2009); and texts at the B2 level from three Reading papers in the FCE Handbook for teachers (Cambridge ESOL 2007) and Top Tips for FCE (Cambridge ESOL 2008).

To automatically measure textual features (see Table 9) of the GEPT and Cambridge English test papers, Coh-Metrix version 2.1 (Graesser, McNamara, Louwerse and Cai 2004, McNamara, Louwerse and Graesser 2002), VocabProfile version 6.2 (Cobb 2010), and WordSmith version 5.0 (Scott 2009) were employed in this study:

1. Coh-Metrix, a free online software tool which incorporates theories of text processing, cognitive psychology, and computational linguistics, is 'sensitive to cohesion relations, world knowledge, and language and discourse characteristics' (Graesser et al 2004). Therefore, unlike

Table 9 Contextual parameters and instruments used in the automatic textual analysis

Contextual parameter		Definition	Instrument
Text length		The average number of words per text	WordSmith
Lexical complexity	Characters/word	The average number of characters per word	WordSmith
	1k word frequency	The ratio of words that appear in the first most frequent 1,000 BNC (2001) wordlist to the total number of words per text	VocabProfile
	2k word frequency	The ratio of words that appear in the second most frequent 1,000 BNC wordlist to the total number of words per text	VocabProfile
	1k+2k word frequency	The ratio of words that appear in the most frequent 2,000 BNC wordlist to the total number of words per text	VocabProfile
	Academic Wordlist (AWL) frequency	The ratio of words that appear in the AWL (Coxhead 1998, 2000) to the total number of words per text	VocabProfile
	Off-list words	The ratio of words that do not appear in either BNC or AWL wordlists to the total number of words per text	VocabProfile
	Celex, raw, mean for content words (0–1,000,000)	The mean raw frequency of all of the content words in a text; the frequency counts come from the CELEX Lexical Database (Baayen, Piepenbrock and Gulikers 1995) A word with the lowest frequency score is the rarest word in the sentence	Coh-Metrix
	Celex, logarithm, mean for content words (0–6)	The log frequency of all content words in a text A word with the lowest log frequency score is the rarest word in the sentence	Coh-Metrix
	Type-token ratio (TTR)	Number of unique words (i.e. types) divided by tokens (i.e. word instances) per text	WordSmith
	Standardised type-token ratio (STTR) basis: 100 words)	Calculated based on every 100 words in order to compare TTR on a common basis	WordSmith

Table 9 (continued)

Contextual parameter		Definition	Instrument
Lexical complexity cont.			
	Lexical density	The ratio of the incidence of lexical words, i.e. nouns, lexical verbs, adjectives and adverbs, to total number of words multiplied by 100	VocabProfile
Syntactic complexity	Average number of words/sentence	The mean number of words per sentence	WordSmith
	Noun Phrase Incidence Score (per 1,000 words)	1,000 divided by the incidence of noun phrase constituents	Coh-Metrix
	Modifiers per noun phrase	The mean number of modifiers, i.e. adjectives, adverbs, or determiners that modify the head noun, per noun phrase Sentences with difficult syntactic compositions have a higher ratio of modifiers	Coh-Metrix
	Higher level constituents	The mean number of higher level constituents per sentence, controlling for number of words Sentences with difficult syntactic composition are structurally embedded and, therefore, have a higher ratio of high-level constituents per word	Coh-Metrix
	Words before main verb	The mean number of words before the main verb of the main clause per sentence Sentences with a larger number of words before the main verb tend to be more difficult	Coh-Metrix
	Sentence syntax similarity, adjacent	The proportion of intersection tree nodes between all adjacent sentences	Coh-Metrix
	Sentence syntax similarity, all, across paragraphs	The proportion of intersection tree nodes between all sentences and across paragraphs	Coh-Metrix

Table 9 (continued)

	Lexical density		The ratio of the incidence of lexical words, i.e. nouns, lexical verbs, adjectives and adverbs, to total number of words multiplied by 100	VocabProfile
Syntactic complexity cont.	Logical operator incidence score		The incidence of logical operators, including ‘and’, ‘or’, ‘not’, ‘if’, ‘then’ and a small number of other similar cognate terms Texts with a high density of these logical operators tend to be more difficult.	Coh-Metrix
Readability	Flesch Reading Ease Score (0–100)		= 206.835 - (1.015 × ASL [average sentence length]) - (84.6 × ASW [average number of syllables per word]) A higher score indicates easier reading. The average document has a Flesch Reading Ease Score between 6 and 70.	Coh-Metrix
	Flesch-Kincaid Grade Level (0–12)		= (.39 × ASL) + (11.8 × ASW) - 15.59 The higher the number, the harder it is to read the text; Flesch Reading Ease Score converted to a USA grade-school level, ranging from 0 to 12	Coh-Metrix
	Coh-Metrix readability		= (Variable 52 content word overlap × 52.230)+ (Variable 49 syntax structure similarity adjacent × 61.306) + (Variable 40 log frequency content words × 22.205) – 45.032 (Crossley et al 2008:487) A higher score indicates more cohesive and easier reading	Manually calculated
Cohesion	Referential cohesion	Ratio of pronouns to noun phrases	The ratio of words classified as pronouns to the incidence of NPs per text A high density of pronouns may create referential cohesion problems when the reader does not know what the pronouns refer to	Coh-Metrix
		Anaphor reference, adjacent, unweighted	The proportion of anaphor references between adjacent sentences Anaphor refers to a word or phrase that relates to other text. A higher score indicates more cohesive and easier reading	Coh-Metrix
		Anaphor reference, all distances, unweighted	The proportion of unweighted anaphor references that refer back to a constituent up to five sentences earlier A higher score indicates more cohesive and easier reading	Coh-Metrix

Table 9 (continued)

Contextual parameter			Definition	Instrument
Cohesion	Referential cohesion	Argument overlap, adjacent, unweighted	The proportion of adjacent sentences that share one or more arguments (i.e. noun, pronoun, NP) or have a similar morphological stem Argument overlap is overlap between the noun in the target sentence and the same noun in singular or plural form in the previous sentence A higher score indicates more cohesive and easier reading	Coh-Metrix
		Argument overlap, all distances, unweighted	The proportion of all sentence pairs per paragraph that share one or more arguments (i.e. noun, pronoun, and noun-phrase) or have a similar morphological stem A higher score indicates more cohesive and easier reading	Coh-Metrix
		Stem overlap, adjacent, unweighted	The proportion of adjacent sentences that share one or more word stems Stem overlap is overlap from the noun to stems, regardless of word type (e.g. noun, verb, and adjective). Both argument and stem overlaps also include overlap between a pronoun and the same pronoun A higher score indicates more cohesive and easier reading	Coh-Metrix
		Stem overlap, all distances, unweighted	The proportion of all sentence pairs per paragraph that share one or more word stems A higher score indicates more cohesive and easier reading	Coh-Metrix
		Proportion of content words that overlap between adjacent sentences	The proportion of content words in adjacent sentences that share common content words A higher score indicates more cohesive and easier reading	Coh-Metrix

Table 9 (continued)

Contextual parameter			**Definition**	**Instrument**
Cohesion cont.	Conceptual cohesion	Latent Semantic Analysis (LSA) adjacent sentences	Mean LSA cosines for adjacent, sentence-to-sentence, units; this measures how conceptually similar each sentence is to the next sentence Text cohesion is assumed to increase as a function of higher cosine scores between text constituents	Coh-Metrix
		LSA all sentences	LSA all sentences computes how conceptually similar each sentence is to every other sentence in the text; all sentence combinations are considered	Coh-Metrix
Text abstractness	Concreteness, mean for content words		The mean concreteness value of all content words per text: match words in the MRC Psycholinguistics Database (Coltheart 1981), which contains 150,837 words and provides information of up to 26 different linguistic properties of these words The higher the number, the more concrete the text is	Coh-Metrix
	Concreteness, minimum in sentence for content words		The mean of the low-concreteness words across sentences. For each sentence in the text, a content word that has the lowest concreteness rating is identified The higher the number, the more concrete the text is	Coh-Metrix
	Mean hypernym values of nouns		The mean hypernym value of nouns in the text: a hypernym metric is the number of levels per conceptual taxonomic hierarchy above (superordinate to) a word A word having more hypernym levels is more concrete. A word with fewer hypernym levels is more abstract	Coh-Metrix
	Mean hypernym values of verbs		The mean hypernym value of main verbs in the text A word having more hypernym levels is more concrete	Coh-Metrix

Note: Adapted from Green, Ünaldi and Weir (2010) and Coh-Metrix users' manual (cohmetrix.memphis.edu/CohMetrixWeb2/HelpFile2.htm) traditional text readability formulas which measure text difficulty solely by word length and sentence length, Coh-Metrix can quantitatively reflect a wide range of aspects of language, in terms of lexical complexity, structural complexity, cohesion, and text abstractness.

2. VocabProfile, also a free online software tool, provided information about lexical complexity, such as the percentage of words occurring among the most frequent and the second most frequent 1,000 words in the BNC (BNC Consortium 2001), the percentage of words in a text appearing in Academic Wordlist (Coxhead 1998, 2000), and lexical density (number of content words as a proportion of the number of grammatical words).
3. WordSmith was used to provide information about lexical and syntactic complexity, such as the average number of characters per word, the average number of words per sentence, the total number of words in a text, and the ratio of different words to tokens.

For details, see Table 9.

Qualitative, non-automated analysis of remaining contextual parameters

Textual characteristics that were not measurable by the automated tools were analysed through expert judgement using GEPT Intermediate Level Past Paper–5 (Language Training and Testing Center 2011a) targeting B1 level and High-Intermediate Level Past Paper–5 (Language Training and Testing Center 2011b) targeting B2 level; and Paper 1 in the Cambridge PET Handbook (Cambridge ESOL 2009) targeting B1 level and Paper 1 in the FCE Handbook (Cambridge ESOL 2007) targeting B2 level.

A Contextual Parameter Proforma (see Table 10 and Appendix 2) was developed based on three tables relevant to the reading comprehension that the Manual (Council of Europe 2009) provided: i.e. Form A10 (Reading Comprehension), Form A19 (Aspects of Language Competence in Reception), and the CEFR Content Analysis Grid for Reading (Alderson et al 2006), supplemented with contextual features identified in Khalifa and Weir's (2009) framework for the validation. Expert judges examined various contextual features of the tests under review and used the Proforma to present an overview of the tests and to quantify their judgement on both text and item dimensions of the test tasks. Any criterial distinctions between different tests could then be identified.

Table 10 Contextual Parameter Proforma

In the following tasks, you will evaluate (names of the exams) texts and accompanying items. Think about any criterial differences between the (name of the exam) tasks and the (name of the exam) tasks for later report back to the whole group in the workshop.

Contextual parameters	**Name of the exam (CEFR level)**
Test setting	
General purpose	
Target population	
Structure of the test	
Test focus (general levels of proficiency the test intends to cover, along with a description of the particular subskills to be tested)	
Communicative topics	
Authenticity	
Time constraint	
Overall number of words	
Number of texts	
Maximum number of words for any single text	
Expected speed of reading	
Task setting	
Item dimension	
Answer type	
Scoring method	
Scoring criteria	
Weighting	

Table 10 (continued)

In the Appendix, you will find a set of reading paper from the (name of the exam). Please refer to the texts and accompanying items for this task. Then answer each of the questions about the text and items by circling the appropriate response.

Contextual parameters			**Name of the exam (CEFR level)**				
Text dimension			**Part I, Questions 1–2**				
Domain		I–1	① Social	② Work		③ Academic	
Discourse mode	Genre	I–2	① Public sign/notice	② Advertisement/leaflet/brochure		③ Letter/memo/email message	
			④ Magazine and newspaper article/report			⑤Fiction book	
	Rhetorical task	I–3	① Exposition			②Argumentation/persuasion/evaluation	
			③ Historical biographical/autobiographical narrative				
Rhetorical organisation		I–4	The organisational structure of the text is				
			① Explicit	②	③	④	⑤ Not explicit
Subject specificity		I–5	Is the topic of the text of general interest or does it require subject-specific knowledge on the part of the reader?				
			① General	②	③	④	⑤ Specific
Cultural specificity		I–6	Is the topic of the text culture neutral or is it loaded with specific cultural content?				
			① Culture neutral	②	③	④	⑤ Culture specific
Text abstractness		I–7	Is the text concrete or abstract?				
			① Concrete	②	③	④	⑤ Abstract
Item dimension							
1	Content dimension	1–1	① Main idea	② Detail	1–2	① Fact	② Opinion
	Explicitness dimension	1–3	① From explicit information		② From implicit information		
	Did you find the information to answer the question _______?			1–4	① Within sentence	② Across sentences	③ At the whole text level
2	Content dimension	2–1	① Main idea	② Detail	2–2	① Fact	② Opinion
	Explicitness dimension	2–3	① From explicit information	② From implicit information			
	Did you find the information to answer the question _______?			2–4	① Within sentence	② Across sentences	③ At the whole text level

Feedback on methodology used to analyse contextual parameters

To investigate whether the judges considered the results from the automated analysis of the textual features useful in determining test levels, Feedback Evaluation Questionnaire–1 (see Table 11) was developed. The more useful the judges felt the indexes generated by the automated tools, the more likely that the indexes could reliably reflect human judgements on textual characteristics.

Feedback Evaluation Questionnaire–2 (see Table 12) was devised for judges to reflect how confident they felt when using the Contextual Parameter Proforma to make judgements. The more confident the judges felt, the more likely their judgements were reliable.

Procedures for data collection

Twelve judges were trained to analyse textual features of the GEPT and Cambridge English reading tests at the B1 and B2 levels. The following procedure was adapted from the Familiarisation and Specification procedures the Manual (Council of Europe 2003, 2009) suggests:

1. Familiarised themselves with the Contextual Parameter Proforma (see Table 10 and Appendix 2) and discussed with the researcher if they had any questions concerning the Proforma.
2. Responded to the Cambridge PET and FCE tasks as if they were taking the tests and working independently, and applied the Contextual Parameter Proforma to analyse the tasks.
3. Handed in their responses to the Proforma to the researcher for statistical analysis and received the results from the automated textual analysis on Cambridge English Reading tasks at the B1 and B2 levels.
4. Reflected how relevant they considered the textual features in automated analysis when they made holistic judgement on levels of task difficulty.
5. Repeated the same process to analyse GEPT reading papers at the B1 and B2 levels, i.e. one GEPT Intermediate paper and one GEPT High-Intermediate paper.
6. Attended a group session to discuss with the researcher and other judges the results of the analyses of their responses to the Contextual Parameter Proforma and that of the automated textual analyses and explored the extent to which their responses agreed with or differed from the other judges' responses and the automated textual analyses.
7. Answered the Feedback Evaluation Questionnaire–1 (see Table 11) regarding how useful they found the results from the automated textual

Table 11 Feedback Evaluation Questionnaire–1

Based on the workshop experience, how useful do you find the parameters below in determining differences between different test levels, e.g. CEFR B1 and B2 levels?

Please circle 0, 1, or 2 (0 = not useful; 1= of limited use; 2 = useful).

Contextual parameters		I find it _____		
	Text length	0	1	2
Vocabulary	Characters/word	0	1	2
	1k word frequency	0	1	2
	2k word frequency	0	1	2
	1k+2k word frequency	0	1	2
	Academic wordlist (AWL) frequency	0	1	2
	Type-token ratio (TTR)	0	1	2
	Standardised type-token ratio (STTR)	0	1	2
	Lexical density	0	1	2
Grammar	Average number of words/sentence	0	1	2
	Higher level constituents	0	1	2
	Noun Phrase Incidence Score	0	1	2
	Mean number of modifiers per noun phrase	0	1	2
	Mean number of words before main verb	0	1	2
	Logical operator incidence Score	0	1	2

Table 11 (continued)

Based on the workshop experience, how useful do you find the parameters below in determining differences between different test levels, e.g. CEFR B1 and B2 levels?

Please circle 0, 1, or 2 (0 = not useful; 1= of limited use; 2 = useful).

Contextual parameters		I find it _____			
Readability		Flesch Reading Ease Score	0	1	2
		Flesch-Kincaid Grade Level	0	1	2
		Coh-Metrix readability	0	1	2
Cohesion	**Referential cohesion**	Anaphor reference	0	1	2
		Argument overlap	0	1	2
		Content word overlap	0	1	2
	Conceptual cohesion	Latent Semantic Analysis (LSA) adjacent	0	1	2
		LSA all sentences	0	1	2
Text abstractness		Concreteness, mean for content words	0	1	2
		Concreteness, minimum in sentence for content words	0	1	2
		Mean hypernym values of nouns	0	1	2
		Mean hypernym values of main verbs	0	1	2

Table 12 Feedback Evaluation Questionnaire–2

Based on the experience applying the Contextual Parameter Proforma, how confident did you feel when you chose the response? Please circle A, B, or C. If your answer is C, please specify.

A = Confident

B = Not confident because I was not sure if I was applying the categories appropriately

C = Not confident because I think some important concepts were not addressed and there should be more categories

Parameter	I felt . . .			
Domain (Social, work, or academic)	A	B	C	(additional categories ______________________)
Genre	A	B	C	(additional categories ______________________)
Rhetorical task	A	B	C	(additional categories ______________________)
Rhetorical organisation	A	B	C	(additional categories ______________________)
Subject specificity	A	B	C	(additional categories ______________________)
Cultural specificity	A	B	C	(additional categories ______________________)
Text abstractness	A	B	C	(additional categories ______________________)
Content dimension (Main idea vs detail)	A	B	C	(additional categories ______________________)
Content dimension (Fact vs opinion)	A	B	C	(additional categories ______________________)
Explicitness dimension (from explicit information vs from implicit information)	A	B	C	(additional categories ______________________)
Did you find the information to answer the question within sentence, across sentences, or at the whole text level?	A	B	C	(additional categories ______________________)

analysis in differentiating task difficulty at different test levels and the Feedback Evaluation Questionnaire–2 (see Table 12) concerning how confident they felt when they made judgement using the Contextual Parameter Proforma.

Data analysis

The Mann-Whitney test, a non-parametric significance test, using SPSS version 16 (2007), was performed to determine the significance of the observed differences between (1) the GEPT tasks at the B1 and B2 levels, (2) the Cambridge English tasks at the B1 and B2 levels, (3) the GEPT and the Cambridge English tasks at the B1 level, and also (4) the GEPT and the Cambridge English tasks at the B2 level in the results from the automated textual analysis. These comparisons involved few data points and the data were not normally distributed. The Mann-Whitney test was therefore selected for the analysis as it does not require a large sample size and a normal distribution of data is not necessary. The Mann-Whitney test compares the medians of two groups of ordinal, non-parametric data to determine if they are statistically different. For those textual characteristics that are not measurable by automated tools, descriptive statistics were computed through SPSS based on the qualitative judgement. Frequencies and modes for items on nominal scales and means for items on five-point Likert scales were computed.

Finally we turn to the cognitive validity parameters of both sets of tests GEPT and Cambridge English at the B1 and B2 levels. Procedures to generate evidence on this validity component will be described to enable us to compare GEPT examinations with equivalent Cambridge English examinations across and between levels.

Cognitive processing analysis: Vertical and horizontal comparisons of GEPT and Cambridge English exams at the B1 and B2 levels

The cognitive processing involved will contribute to the difficulty of completing tasks in a test. Within a GEPT test we would expect the cognitive processing demands to be higher in a B2 level examination than in a B1 level examination. Similarly when we compare two tests (Cambridge English and GEPT) deemed to be at a comparable level, we would expect a good degree of similarity between the cognitive processing demands made in each. This section describes the methodology for establishing evidence in respect of the cognitive processing required in both the GEPT and Cambridge English tasks.

Instruments

The reading paper of GEPT Intermediate Past Paper–5 (Language Training and Testing Center 2011a), the reading paper of GEPT High-Intermediate Past Paper–5 (Language Training and Testing Center 2011b), Paper 1 in Cambridge PET Handbook for teachers (Cambridge ESOL 2009), and Paper 1 in Cambridge FCE Handbook for teachers (Cambridge ESOL 2007) were used to collect data from both judges and test takers.

To investigate what the test is designed to assess from the experts' perspective, the Cognitive Processing Proforma (see Table 13 and Appendix 3) was designed based on Khalifa and Weir's (2009:43) framework to quantify expert judgement on what cognitive processes were involved when test takers were taking the GEPT and Cambridge English reading tests at the B1 and B2 levels. To evaluate how confident judges felt when making judgements on what cognitive processes took place when test takers were taking the reading tests, the Cognitive Processing Feedback Evaluation Questionnaire (see Table 14) was developed.

Table 13 Cognitive Processing Proforma

In the Appendix you will find a set of reading paper from (name of the exams). Please indicate by a tick that you think a particular cognitive process takes place when test takers are answering the questions in each part of the tests. Think about any criterial differences between these two examinations for later report back to the whole group in the workshop.

Cognitive processing	Name of the exam (B1)					Name of the exam (B2)		
	Part 1	Part 2	Part 3	Part 4	Part 5	Part 1	Part 2	Part 3
Word recognition								
Lexical access								
Syntactic parsing								
Establishing propositional meaning at clause and sentence level								
Inferencing								
Integrating information across sentences								
Creating a text level structure								
Integrating information across texts								

Table 14 Cognitive Processing Feedback Evaluation Questionnaire

Based on the experience applying the Cognitive Processing Proforma, how confident did you feel when you chose your response? Please circle A, B, or C. If your answer is C, please specify.

A = Confident

B = Not confident because I am not sure if I was applying the categories appropriately

C = Not confident because I think some important concepts were not addressed

Parameter	I felt . . .		
Word recognition	A	B	C
Lexical access	A	B	C
Syntactic parsing	A	B	C
Establishing propositional meaning at clause and sentence level	A	B	C
Inferencing	A	B	C
Integrating information across sentences	A	B	C
Creating a text level structure	A	B	C
Integrating information across texts	A	B	C

To investigate what cognitive processing skills the test takers were using from the test takers' perspective, the Cognitive Processing Checklist (henceforward 'the Checklist'; see Table 15) was designed, based on categorisation of reading types (Urquhart and Weir 1998:123), for test takers to report what they actually did to find the answers to each test question.

The Checklist, consisting of eight items for test takers to report what they actually did to find the answer to each test question, was piloted on 81 target test takers of the GEPT Intermediate level, randomly divided into two groups: Group 1, 39 test takers in total, and Group 2, 42 test takers. Both groups took the GEPT Intermediate listening and reading tests. After the listening and reading tests, they were instructed to practice responding to the eight questions on the Checklist immediately after they answered each question to report what cognitive processing skills they actually used when solving a reading task. After they familiarised themselves with the Checklist, Group 1 was given a GEPT Intermediate multiple-choice gap-filling task with five blanks, and Group 2 a GEPT Intermediate reading text followed by four multiple-choice comprehension questions. Both groups were given a maximum of 10 minutes to answer the reading task and responded to the checklist. Most of the test takers in Group 1 completed the reading task and the Checklist in 5 to 6 minutes, and Group 2 within 5 to 7 minutes. In order to compare cognitive processing skills that the passing and non-passing test takers used, both groups were further divided into two, i.e. the Passing group

Table 15 Cognitive Processing Checklist (Pilot version)

Directions: immediately after answering each question, read each of the statements below and indicate by a tick any that match what you did

To find the answer to the question, I tried to . . .		1	2	3	4
i.	quickly match words that appear in the question with similar or related words in the text	Y N	Y N	Y N	Y N
ii.	search quickly for relevant part(s) of the text and read them carefully	Y N	Y N	Y N	Y N
iii.	connect information from the text with knowledge I already have	Y N	Y N	Y N	Y N
iv.	understand ideas which are not explicitly stated	Y N	Y N	Y N	Y N
v.	read the whole text slowly and carefully	Y N	Y N	Y N	Y N
I found the answer . . .		1	2	3	4
vi	within a single sentence	Y N	Y N	Y N	Y N
vii	by putting information together across sentences	Y N	Y N	Y N	Y N
viii	by understanding how information in the whole text fits together	Y N	Y N	Y N	Y N

(18 test takers from Group 1 and 15 from Group 2) and the Non-passing group (21 test takers from Group 1 and 27 test takers from Group 2), based on their scores from the GEPT Intermediate tests.

In this study, cognitive skills, i.e. the independent variable, was nominal and the results might not meet the assumption of normality, a prerequisite for parametric correlation analyses. Therefore, Spearman rank correlation, a non-parametric form of correlation, was performed to investigate the relationships between cognitive processing skills that passing and failing test takers used.

The pilot results (see Table 16) showed that test results were significantly correlated to their responses to the Checklist on six out of the nine test questions (highlighted in bold in the table). In addition, when responding to the gap-filling task, the Passing test takers used scanning skills significantly more often than the Non-passing ones ($p<.05$), while the Non-passing test takers employed careful reading significantly more often than the Passing test takers ($p<.01$).

Table 16 Cognitive Processing Checklist pretest statistics

	Task type	Multiple-choice gap-filling passage with five blanks						Text followed by four multiple-choice comprehension questions				
	Test focus	**Local lexis**	**Global lexis**	**Global syntax (cohesive device)**	**Global syntax (tense)**	**Local lexis**	**P=18 N=21**	**Main idea**	**Detail**	**detail**	**Infer.**	**P=15 N=27**
	Statistics	**Spearman correlation**					**t-test**	**Spearman correlation**				**t-test**
To find the answer to the question, I tried to . . .	**Question no.** **Cognitive operations**	**1**	**2**	**3**	**4**	**5**		**1**	**2**	**3**	**4**	
i. quickly match words that appeared in the question with similar or related words in the text.	**Scanning**	.18	.19	−.16	.00	.12	2.03 **(.05*)**	.12	.14	.04	−.20	.93 (.36)
ii. search quickly for relevant part(s) of the text and read them carefully.	**Search reading**	.07	.07	−.23	.01	.30	.85 (.40)	.18	**.37***	**−.33***	−.08	.48 (.63)
iii. connect information from the text with knowledge I already have.	**Inferencing**	.02	−.09	−.21	.07	−.19	−.68 (.50)	−.20	**−.31***	−.06	−.08	1.02 (.31)
iv. understand ideas which are not explicitly stated.	**Inferencing**	.06	**.35***	−.16	**.30***	−.04	.36 (.72)	.22	−.05	.05	.18	−1.37 (.18)
v. read the whole text slowly and carefully.	**Careful reading**	−.20	−.10	−.18	−.18	**−.45***	−2.60 **(.01**)**	−.04	−.05	.11	.24	−.67 (.51)

Table 16 (continued)

		Task type	**Multiple-choice gap-filling passage with five blanks**						**Text followed by four multiple-choice comprehension questions**				
		Test focus	**Local lexis**	**Global lexis**	**Global syntax (cohesive device)**	**Global syntax (tense)**	**Local lexis**	**P=18 N=21**	**Main idea**	**Detail**	**detail**	**Infer.**	**P=15 N=27**
		Statistics	**Spearman correlation**					**t-test**	**Spearman correlation**				**t-test**
I found the answer . . .		**Question no.** **Level of comprehension**	1	2	3	4	5	t-test	1	2	3	4	t-test
vi	within a single sentence.	**Intra-sentential understanding**	**.33***	.09	−.11	−.25	.13	.64 (.52)	−.07	−.01	**.41****	−.04	−1.52 (.14)
vii	by putting information together across sentences.	**Inter-sentential understanding**	**−.40***	−.15	.01	.15	−.18	−1.97 (.06)	.10	−.05	**−.37***	.25	.57 (.57)
viii	by understanding how information in the whole text fits together.	**Text-level understanding**	−.22	−.12	.07	.23	−.15	−1.33 (.19)	.07	−.03	.07	−.03	−.32 (.75)

Note: P = Passing test takers; N = Non-passing test takers
**p <.05; **p <.01*

After examining the test takers' responses to the Checklist more closely, it was found that the format of the Checklist might have been misleading to some test takers. When asked how they found their answer to a reading question, some test takers selected more than one 'Yes' to the Checklist items 6 to 8; e.g. they reported finding the answer both within a sentence and across sentences. To ensure that the test takers reported the two operations separately as intended, the eight items were then re-categorised into two items; the original items 1 to 5 were re-coded as five choices under Item 1 'To find the answer to the question, I tried to . . .', and the original items 6 to 8 were re-coded as three choices under Item 2. In addition, the test takers were explicitly informed that it was possible to choose more than one option from Item 1, while they should choose only one option from Item 2 (see Table 17 and Appendix 1).

Procedures for data collection

The same judges responding to the Contextual Parameter Proforma (see Table 10 and Appendix 2) were requested to fill out the Cognitive Processing Proforma (see Table 13 and Appendix 3) immediately after they analysed the same test using the Contextual Parameter Proforma. The procedure is described below:

1. Familiarised themselves with the Cognitive Processing Proforma (see Table 13 and Appendix 3) and discussed with the researcher if they had any questions concerning the Proforma.
2. Responded to the Cambridge PET and FCE tasks as if they were taking the tests, and, working independently, applied the Cognitive Processing Proforma to analyse the tasks.
3. Handed in their responses to the Proforma to the researcher for statistical analysis.
4. Repeated the same process to analyse GEPT reading papers at the B1 and B2 levels, i.e. one GEPT Intermediate paper and one GEPT High-Intermediate paper.
5. Attended a group session to discuss with the researcher and other judges the results of the analyses of their responses to the Cognitive Processing Proforma and explored the extent to which their responses agreed with or differed from the other judges' responses.
6. Answered the Cognitive Processing Feedback Evaluation Questionnaire (see Table 14) concerning how confident they felt when they made judgements using the Cognitive Processing Proforma.

To investigate the cognitive processing skills used from a test takers' perspective, 71 target test takers of the GEPT Intermediate level, targeting CEFR B1 level, and 73 target test takers of the GEPT High-Intermediate level, targeting CEFR B2 level, participated in the study; see Table 18 for data collection design (see also Table 8). Using the Cognitive Processing Checklist (see Table 17 and Appendix 1), they reported what they had actually done to find the answer when responding to each question.

Table 17 Cognitive Processing Checklist (for the main study)

Directions: Immediately after you answer each question in this reading test, please indicate what you actually did to find the answer to the question by ticking the appropriate choice(s) to checklist questions 1 and 2.

					Question 1 (one or more answers to be chosen)					**Question 2 (only one answer to be selected)**		
	Answer to the reading test				To find the answer to the question, I tried to . . . A quickly match words that appeared in the question with similar or related words in the text. B search quickly for part(s) of the text which might answer the question and read them carefully. C connect information from the text with knowledge I already have. D understand ideas which are not explicitly stated. E read the whole text slowly and carefully to find the answer to the question.					I found the answer . . . A within a single sentence. B by putting information together across sentences. C by understanding how information in the whole text fits together.		
1.	A	B	C	D	A	B	C	D	E	A	B	C
2.	A	B	C	D	A	B	C	D	E	A	B	C
3.	A	B	C	D	A	B	C	D	E	A	B	C
.												
.												
.												

A single group counter-balanced design was used. The GEPT Intermediate level target examinees took both the GEPT and Cambridge English tests and were randomly divided into two groups (see Table 18, and also Table 8): Groups 3 and 4 took the GEPT and Cambridge English reading tests at the B1 level in two consecutive sessions, with a 10-minute break in between, and were asked to fill out the Checklist immediately after they answered

Table 18 Cognitive processing data collection design (see also Table 8)

	B1 Level (N=71)		B2 Level (N=73)	
	Group 3 (N=36)	**Group 4 (N=35)**	**Group 3 (N=31)**	**Group 4 (N=42)**
Test Session 1	GEPT Intermediate and the checklist in 55 minutes	Cambridge PET and the checklist in 60 minutes	GEPT High-Intermediate and the checklist in 60 minutes	Cambridge FCE and the checklist in 75 minutes
Break	10 minutes	10 minutes	10 minutes	10 minutes
Test Session 2	Cambridge PET and the checklist in 60 minutes	GEPT Intermediate and the checklist in 55 minutes	Cambridge FCE and the checklist in 75 minutes	GEPT High-Intermediate and the checklist in 60 minutes

each comprehension question, and the GEPT High-Intermediate level target examinees underwent the same process. Those who took the GEPT Intermediate, the GEPT High-Intermediate or Cambridge PET were given an extra 10 minutes, and those who took the FCE 15 minutes, based on the Checklist pre-test result presented earlier in this chapter, to compensate for the time they spent responding to the Checklist. To minimise any practice effect, the order of administering the GEPT and Cambridge English tests was counterbalanced: Group 3 took the GEPT and answered the Checklist first and then took the Cambridge English test and answered the Checklist, while Group 4 took the Cambridge English test and answered the Checklist first and then the GEPT and responded to the Checklist.

Data analysis

The data analysis involved information obtained from two sources:

1. Expert judgement collected through the Cognitive Processing Proforma (see Table 13 and Appendix 3) and
2. Test takers' self-report using the Cognitive Processing Checklist (see Table 17 and Appendix 1).

The frequencies of the eight cognitive skills that experts assumed the test takers used when they responded to each reading task were counted and weighted based on the tasks' contribution to the total score of the test in question, and then averaged so that tasks with different numbers of test questions could be compared on a common basis.

Before data from the test takers' self-report on their use of cognitive processing skills were analysed, it was important to investigate whether the test takers' performance on the GEPT and Cambridge English reading tests was affected by the order of administration, as well as whether their performance was affected when they were asked to respond to the checklist. ANOVA was performed on the test takers' scores on the GEPT and Cambridge English tests at the same CEFR level. If no significant difference was observed on performance among the four groups, which showed that test takers' performance was not significantly affected by the order of administration nor the administration of the Checklist together with the exams ($p<.05$), scores from all tests were then pooled together so that the analysis could be performed based on a larger sample. The result then suggested that those who took the reading tests and also responded to the Checklist went through the same cognitive processes as those who took the reading tests only.

Based on their test scores, test takers were rank-ordered: those who scored the highest 27% on the GEPT tests and also the highest 27% on the Cambridge English tests were identified as the High Group, i.e. those with high English reading ability, and those who scored the lowest 27% on the GEPT tests and also the lowest 27% on the Cambridge English tests were identified as the Low Group, i.e. those with low English reading ability (based on Henning 1987). Means, frequencies, and standard deviations of test takers' responses to the Checklist were calculated separately on three groups: the High Group, the Low Group and the Whole Group. Since several assumptions of one-way ANOVA (equal interval and a normal distribution of the data) were not met to detect differences in test takers' responses to the Checklist, the Friedman test, a non-parametric ANOVA alternative, was performed to compare whether differences in cognitive operations that the High Group, the Low Group, and the Whole Group reached significance.

Within-group differences that reached significance based on the Friedman test were tested using the Wilcoxon signed-ranks test, a non-parametric alternative to the paired t-test, to learn whether the examinees in the same group, i.e. the High Group, the Low Group, or the Whole Group, processed the reading tasks differently. According to the Dunn-Bonferroni correction, the alpha level was adjusted to 0.125 for Item 1, since the comparison was carried out four times; and for Item 2, the alpha level was adjusted to 0.25 since the comparison was carried out twice. Afterwards, the Mann-Whitney test was performed to investigate whether the High and the Low groups processed the reading tasks differently, and also to examine whether test takers used different cognitive skills when they were taking the GEPT and Cambridge English reading tests at the same CEFR level.

Results on vertical comparisons of the GEPT at CEFR B1 and B2 levels, in terms of test results, and contextual and cognitive processing parameters, will be reported in Chapter 4 to answer Research Question 1: Is a GEPT reading test designed to measure at CEFR B2 level more difficult than a GEPT reading test designed to measure at CEFR B1 level in terms of test results, contextual parameters, and cognitive processing skills? Results on horizontal comparisons between the GEPT and Cambridge English reading tests at the same CEFR level, in terms of test results, and contextual and cognitive processing parameters, will be reported in Chapter 5 to answer Research Question 2: Are GEPT reading tests at CEFR B1 and B2 levels comparable to alternative CEFR-linked measures in terms of test results, contextual parameters, and cognitive processing skills?

4 Results and discussion 1: Vertical comparisons of GEPT reading tests at CEFR B1 and B2 levels

Chapter overview

This chapter reports results on the validation of the GEPT level framework in terms of test scores, contextual parameters, and cognitive processing involved. This addresses Research Question 1: Is a GEPT reading test designed to measure at CEFR B2 level more difficult than a GEPT reading test designed to measure at CEFR B1 level in terms of test results, contextual parameters, and cognitive processing skills? Results from vertical scaling scores from different levels of the GEPT onto a common score scale are reported, followed by results from analysis of contextual parameters that affect the comprehensibility and difficulty of reading tasks and those of cognitive processing skills involved when examinees are taking the GEPT at CEFR B1 and B2 levels, respectively.

Vertical scaling

Scores from different levels of a multilevel test are based on different score scales, and score units of these tests are not necessarily the same. Therefore, they cannot be compared directly. To empirically investigate whether test difficulty increases as the GEPT level advances, scores from different levels of the GEPT reading tests were linked onto a common score scale to allow comparisons of the degrees of difficulty across the GEPT levels in this study. The scope of this study was limited to CEFR B1 and B2 levels; nevertheless, to provide a more complete picture, the current study linked three dichotomously scored GEPT tests, namely the Elementary, Intermediate, and High-Intermediate reading tests, onto the scale. See Table 4 in Chapter 3 for an overview of the test formats of the GEPT Elementary, Intermediate, and High-Intermediate reading tests.

Data collection design

The common-item non-equivalent groups design was adopted to link scores from different levels of the GEPT onto a common score scale. A total of 827 students took part in the study: 429 14-to 17-year-old students, target examinees of Elementary and Intermediate levels, were invited from three high schools to take Testlet 1, and 398 17-to 21-year-old students, target examinees of Intermediate and High-Intermediate levels, were invited from one high school and three colleges to take Testlet 2. For information on the design of the data collection method used in this study, see Table 19.

Table 19 Vertical scaling data collection design

Examinee group	**Testlet by GEPT level**		
	Elementary	**Intermediate**	**High-Intermediate**
Group 1 (N=429)	Testlet 1		
Group 2 (N=398)		Testlet 2	

Results based on CTT analyses

Cronbach's coefficient alphas for Testlet 1 and Testlet 2 (0.83 and 0.88, respectively; see the shaded area in Table 20) suggested that the reliability of the two testlets conformed to commonly acceptable standards for large-scale exams. The means and standard deviations of the number of correct items were 22.87 and 6.76, respectively, for Testlet 1, and 32.09 and 7.54, respectively, for Testlet 2. The percentage correct of the total items, i.e. 0.57 and 0.71 (see the shaded area in Table 20), respectively, indicated that the difficulty of Testlet 1 for Group 1 examinees was appropriate, while Testlet 2 was relatively easy for Group 2 examinees. The percentages correct of the common items of Group 1 and Group 2 were 0.47 and 0.79 (see the shaded area in Table 20), respectively, and the difference between Groups 1 and 2 was 0.32, indicating that the English proficiency of the two groups of examinees differed. For details, see Table 20.

Results based on IRT analyses

In this study, BILOG-MG (Zimowski et al 2003) with a group option was employed to scale Rasch model item and ability estimates. Both concurrent estimation and separate estimation were performed and compared, although few significant differences between parameters estimated with the two methods were observed in previous research (Camilli et al 1993, Hanson and Beguin 2002). Concurrent calibration provides more accurate estimates

Table 20 Descriptive statistics for GEPT reading Testlets 1 and 2

		Testlet 1	Testlet 2
No. of examinees		429	398
No. of items		40	45
Mean		22.87	32.09
Mean % correct		0.57	0.71
Standard deviation		6.76	7.54
Minimum		6	8
Maximum		40	45
Alpha		0.83	0.88
Common items	No. of items	25	25
	Mean	11.82	19.75
	Mean % correct	0.47	0.79
	Standard deviation	4.35	4.09
	Minimum	3	4
	Maximum	25	25

for the common (Intermediate Level in this study) items since the estimation is based on a larger sample size as responses from both test administrations are included, while separate calibration is more robust to violations of the IRT assumptions due to multidimensionality. In this study, concurrent estimates were used operationally, while separate estimates were computed to triangulate the results from concurrent estimation and to identify potential problems.

Difficulty (*b*) parameter statistics

The means of difficulty parameter (*b*) estimates of the Elementary, Intermediate, and High-Intermediate reading test items were –1.57, –0.01, and 1.25, respectively, based on concurrent estimation, and –1.59, –0.04, and 1.31, respectively, based on separate estimation. The difference between the Elementary and Intermediate levels was around 1.56, which was slightly larger than the difference between the Intermediate and High-Intermediate levels, around 1.26 (for details, see Table 21), suggesting the increases in test difficulty between two adjacent levels of the GEPT were relatively steady across levels. In general, the means of the scaled scores increased with the GEPT levels.

The spreads of the difficulties of the GEPT Elementary and Intermediate reading test items overlapped roughly to the same extent as those of Intermediate and High-Intermediate items based on results from concurrent and separate estimation (see Figures 2 and 3). The Elementary

Table 21 IRT difficulty parameter estimates (*b*) statistics

GEPT Level		Elementary	Intermediate	High-Intermediate	Total
Number of items		15	26	19	60
Concurrent Estimation	Mean *b*	−1.57	−0.01	1.25	0.00
	SD*	0.85	0.65	0.98	1.33
Separate Estimation	Mean *b*	−1.59	−0.04	1.31	0.00
	SD	0.85	0.72	1.05	1.39

**SD = Standard deviation*

Figure 2 Distributions of IRT item difficulty parameter estimates (*b*), based on concurrent estimation by GEPT level

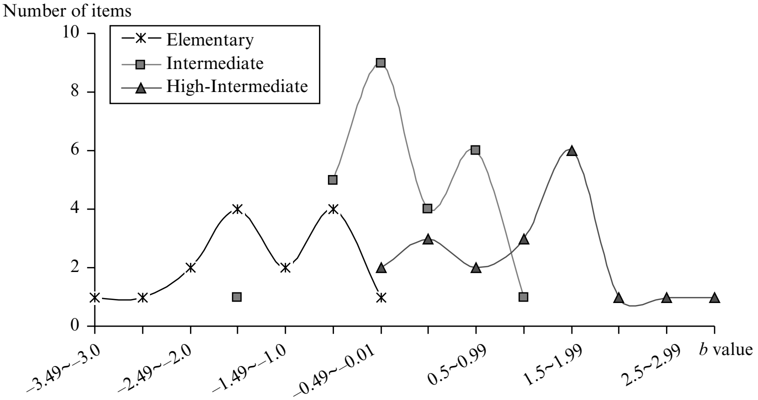
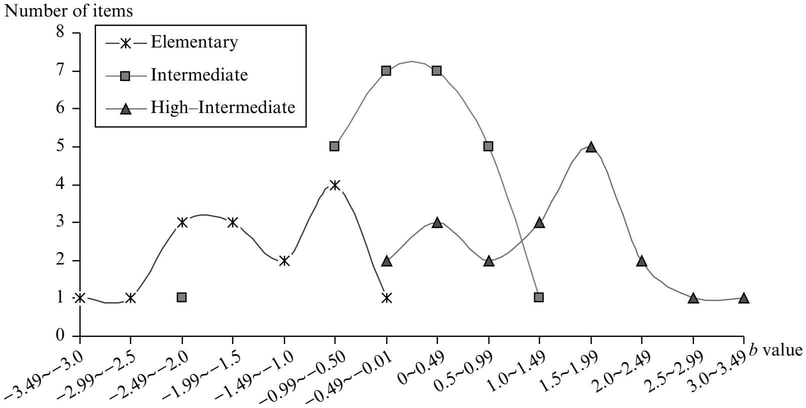


Figure 3 Distributions of IRT item difficulty parameter estimates (*b*), based on separate estimation by GEPT level



and Intermediate curves intersected at –0.5, and the Intermediate and High-Intermediate curves intersected at 1.4.

Differences in *b* parameter estimates and correlations obtained from concurrent and separate estimation of the two reading testlets were compared. The differences in difficulty estimates ranged from –0.36 to 0.41. Based on the results of the paired t-test, there was no significant difference between the estimates obtained using the two estimation methods ($p=0.98$). The correlation between concurrent and separate difficulty estimates was 0.99. The high degree of correlation suggested that the observed data fitted the IRT assumption and the scaling was appropriate. See Table 22.

Ability parameter (θ) statistics

Overall, the distribution of ability parameter estimates (θ) of Group 1 test takers fell in the lower end of the ability estimate axis and those of Group 2 in the upper end of the ability estimate axis, with means of –1.81 for Group 1 and 1.10 for Group 2. See Table 23.

The distributions of the two groups moderately overlapped (see Figure 4), which observed the general requirements that Patz and Yao (2007:253) specified for the examinee population used to produce vertical scales:

> When differences in population proficiency at adjacent levels are modest in comparison to differences between examinees within levels and when the expectations or standards against which examinees are to be measured overlap extensively, then linking the adjacent test levels to a common scale will make sense and provide meaningful information.

Relationships between the CTT test scores from the operational GEPT reading test scores and the IRT ability estimates (θ)

The sample test takers' CTT test scores from the operational GEPT were compared with the IRT ability estimates (θ) in this study in order to investigate the relationships between test takers' performance on the operational GEPT reading tests and their IRT ability estimates. The mean ability estimates for passing candidates of Elementary, Intermediate, and High-Intermediate reading tests were –0.54, 0.65, and 2.40 (see the shaded area in Table 24), respectively; the increase from the Elementary level to the Intermediate level was 1.19, and that from the Intermediate level to the High-Intermediate level was 1.75; see Table 24.

The relationships between the scores obtained from the operational GEPT Elementary and Intermediate reading tests and the increase in the

Table 22 Concurrent and separate item estimates of the reading testlets

Serial No.	GEPT Level	Testlet 1 Item No.	Testlet 2 Item No.	Concurrent estimation	Separate estimation	Difference
R1	Elementary	#3	–	−0.62	−0.65	0.02
R2	Elementary	#6	–	−0.54	−0.56	0.03
R3	Elementary	#8	–	−3.17	−3.18	0.01
R4	Elementary	#10	–	−1.99	−2.01	0.02
R5	Elementary	#12	–	−1.79	−1.81	0.02
R6	Elementary	#16	–	−1.71	−1.73	0.02
R7	Elementary	#17	–	−1.58	−1.60	0.02
R8	Elementary	#18	–	−0.91	−0.93	0.02
R9	Elementary	#19	–	−2.46	−2.47	0.01
R10	Elementary	#20	–	−1.11	−1.13	0.02
R11	Elementary	#26	–	−1.18	−1.20	0.02
R12	Elementary	#27	–	−0.96	−0.98	0.02
R13	Elementary	#31	–	−2.78	−2.79	0.01
R14	Elementary	#32	–	−0.41	−0.44	0.03
R15	Elementary	#33	–	−2.34	−2.36	0.01
R16	Intermediate	#1	#1	−0.10	−0.06	−0.04
R17	Intermediate	#2	#3	0.22	0.27	−0.05
R18	Intermediate	#4	#4	0.72	0.39	0.33
R19	Intermediate	#5	#6	0.59	0.51	0.08
R20	Intermediate	#7	#7	−0.36	−0.45	0.09
R21	Intermediate	#9	#9	0.25	−0.09	0.34
R22	Intermediate	#11	#10	−0.50	−0.75	0.25
R23	Intermediate	#14	#12	−0.09	−0.16	0.07
R24	Intermediate	#13	#13	0.10	0.18	−0.08
R25	Intermediate	#15	#15	−0.13	−0.27	0.14
R26	Intermediate	#21	#16	−0.40	−0.81	0.41
R27	Intermediate	#22	#17	−1.81	−2.21	0.40
R28	Intermediate	#23	#18	−0.60	−0.69	0.09
R29	Intermediate	#24	#19	−0.02	0.34	−0.36
R30	Intermediate	#25	#20	−0.74	−0.84	0.09
R31	Intermediate	#28	#28	1.16	0.97	0.19
R32	Intermediate	#29	#29	−0.30	−0.31	0.01
R33	Intermediate	#30	#30	−0.06	0.19	−0.25
R34	Intermediate	#37	#32	−0.16	0.17	−0.33
R35	Intermediate	#38	#33	0.79	0.55	0.24
R36	Intermediate	#39	#34	0.72	0.71	0.02
R37	Intermediate	#40	#35	0.72	0.94	−0.23
R38	Intermediate	#34	#36	−0.51	−0.34	−0.17

Table 22 (continued)

Serial No.	GEPT Level	Testlet 1 Item No.	Testlet 2 Item No.	Concurrent estimation	Separate estimation	Difference
R39	Intermediate	#35	#37	0.09	0.32	−0.23
R40	Intermediate	#36	#38	0.91	1.07	−0.17
R41	High-Intermediate	–	#2	0.26	0.25	0.01
R42	High-Intermediate	–	#5	1.84	1.95	−0.10
R43	High-Intermediate	–	#8	1.27	1.33	−0.06
R44	High-Intermediate	–	#11	1.52	1.61	−0.08
R45	High-Intermediate	–	#14	1.75	1.84	−0.10
R46	High-Intermediate	–	#21	−0.17	−0.21	0.04
R47	High-Intermediate	–	#22	0.64	0.66	−0.02
R48	High-Intermediate	–	#23	1.83	1.93	−0.10
R49	High-Intermediate	–	#24	1.91	2.02	−0.11
R50	High-Intermediate	–	#25	0.12	0.11	0.02
R51	High-Intermediate	–	#26	−0.38	−0.43	0.05
R52	High-Intermediate	–	#27	2.61	2.76	−0.16
R53	High-Intermediate	–	#31	−0.67	−0.74	0.07
R54	High-Intermediate	–	#39	0.77	0.80	−0.03
R55	High-Intermediate	–	#40	3.23	3.43	−0.20
R56	High-Intermediate	–	#41	1.18	1.24	−0.06
R57	High-Intermediate	–	#42	2.36	2.50	−0.14
R58	High-Intermediate	–	#43	1.15	1.21	−0.06
R59	High-Intermediate	–	#44	0.16	0.15	0.01
R60	High-Intermediate	–	#45	1.68	1.77	−0.09

p=0.98; r=0.99

Table 23 Statistics of ability estimates (θ) by group

Group	No. of students	Mean	SD*
1	429	−1.81	0.20
2	398	1.10	1.27

**SD = Standard deviation*

Figure 4 Distribution of ability estimates (θ) by group

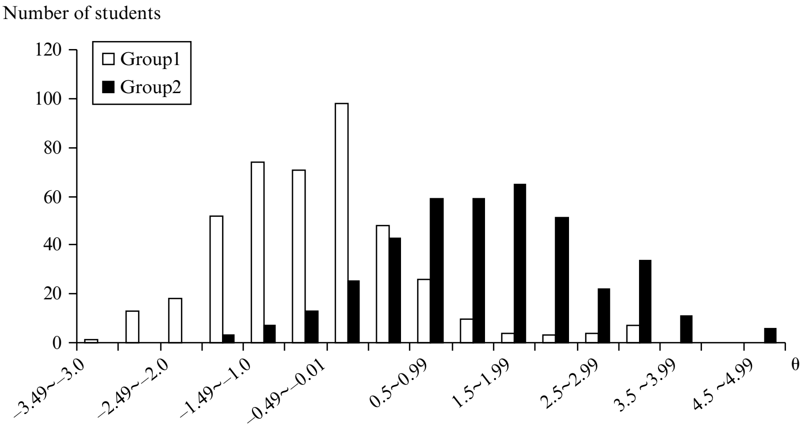
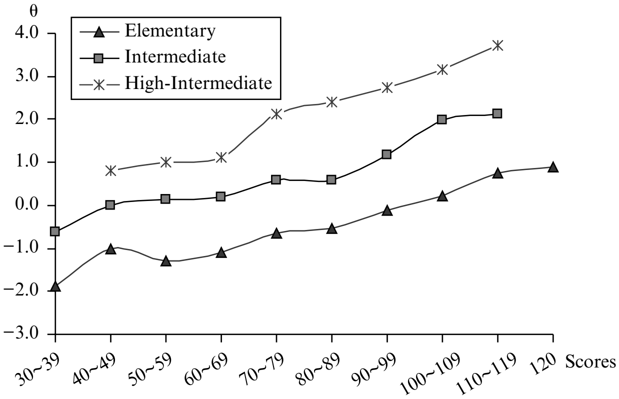


Table 24 The ability estimates (θ) and the GEPT operational reading test scores by GEPT level

Test scores	Elementary	Intermediate	High-Intermediate
120	0.90	–	–
110~119	0.74	2.14	3.71
100~109	0.23	2.00	3.15
90~99	−0.13	1.18	2.74
80~89	−0.54	0.65	2.40
70~79	−0.65	0.59	2.11
60~69	−1.09	0.21	1.12
50~59	−1.29	0.14	1.00
40~49	−1.02	−0.01	0.81
30~39	−1.88	−0.62	–
Number of examinees	201	230	35

ability estimates were relatively linear. The pattern was somewhat irregular for the High-Intermediate level, presumably due to the very small sample size. None of the three curves intersected at any points (see Figure 5).

Figure 5 Distributions of ability estimates (θ) and GEPT operational reading test scores by level



Contextual impacts on reading performance

Textual features, identified in the Chapter 2 literature review, that may affect the comprehensibility and difficulty of reading tasks were analysed using both automated tools and expert judgement. Automated textual analysis was carried out using Coh-Metrix version 2.1 (Graesser et al 2004, McNamara et al 2002), VocabProfile version 6.2 (Cobb 2010), and WordSmith Version 5.0 (Scott 2009). After the indices of each individual text were obtained, they were averaged and tested by the Mann-Whitney U Test, a non-parametric t-test, to determine whether the observed differences between the GEPT reading tests at the B1 and B2 levels reached significance to answer Research Question 1.

As regards the textual characteristics that are not measurable by the automated tools, expert judgement was employed to analyse GEPT tasks at the B1 and B2 levels, using the Contextual Parameter Proforma (see Table 10 and also Appendix 2). The responses to the Proforma were weighted based on the tasks' contribution to the total score of the test in question, and then averaged so that tasks with a different number of test questions could be compared on a common basis.

Results from automated analysis of contextual features

Various aspects of lexical and syntactic complexity, readability, text cohesion, and text abstractness were analysed using automated tools (see Table 9). The indices obtained from the analysis were compared between the GEPT texts at the B1 and B2 levels to determine whether the GEPT texts at the B2 level were significantly more difficult than those at the B1 level. The Mann-Whitney U Test, a non-parametric t-test, was performed between indices obtained based on the GEPT texts at the B1 and B2 levels- GEPT Intermediate (GEPT-I) used texts at Level B1 and GEPT High-Intermediate (GEPT-HI) used text at Level B2 (see Table 25).

Statistically significant differences ($p<.05$) in text length, lexical complexity, average number of words per sentence, readability, and text abstractness were observed (see the shaded areas in Table 25). The GEPT texts at the B2 level used longer texts (see Text length) and contained longer words (see Characters/word) and a wider range of lexis (see 1k+2k word frequency, AWL frequency, and Off-list words (GEPT-1) used texts at Level B1 and GEPT High-Intermediate (GEPT-H1 used text at Level B2)). The sentences were also longer (see Average number of words/sentence). The texts at the B2 level were more difficult, in terms of FREs and FK Grade Level readability, and more abstract (see Concreteness, Mean for content words) than those at the B1 level. Nevertheless, no significant difference in most syntactic complexity and all cohesion indices was observed between the GEPT texts at the B1 and B2 levels.

Table 25 Results on the comparisons between GEPT texts at CEFR B1 and B2 levels based on automated textual analysis

Contextual parameter		**GEPT**		
		GEPT-I	**GEPT-HI**	**Sig.**
Text length		**148.190**	**222.632**	**0.002***
Lexical complexity	Characters/word	**4.596**	**4.929**	**0.001***
	1k word frequency	**77.93%**	**74.14%**	**0.002***
	1k+2k word frequency	**85.78%**	**80.84%**	**0.002***
	AWL frequency	**3.72%**	**5.95%**	**0.008***
	Off-list words	**10.50%**	**13.21%**	**0.036***
	STTR	70.682	72.530	0.258
	Lexical density	0.575	0.571	0.649
Syntactic complexity	Average number of words/sentence	**15.655**	**19.234**	**0.011***
	Noun Phrase Incidence Score (per 1,000 words)	283.404	276.763	0.668
	Mean number of modifiers per noun phrase	0.883	0.928	0.247

Table 25 (continued)

Contextual parameter			**GEPT**		
			GEPT-I	**GEPT-HI**	**Sig.**
Text length			**148.190**	**222.632**	**0.002***
	Mean number of higher level constituents per sentence		0.724	0.715	0.361
	Mean number of words before main verb of main clause in sentences		4.585	4.821	0.789
	Sentence syntax similarity, all, across paragraphs		0.110	0.112	0.872
Readability	Flesch Reading Ease Score		**64.126**	**49.656**	**0.000***
	Flesch-Kincaid Grade Level		**8.114**	**10.471**	**0.000***
	Coh-Metrix readability		14.200	13.133	0.320
Cohesion	Referential cohesion	Ratio of pronouns to noun phrases	0.204	0.146	0.145
		Anaphor reference, all distances, unweighted	0.178	0.108	0.347
		Argument overlap, all distances, unweighted	0.449	0.503	0.555
		Stem overlap, all distances, unweighted	0.421	0.521	0.205
		Proportion of content words that overlap between adjacent sentences	0.075	0.078	0.728
	Conceptual cohesion	Logical operator incidence score (and + if + or + cond + neg)	33.624	35.080	0.452
		LSA adjacent sentences	0.212	0.246	0.178
		LSA all sentences	0.227	0.261	0.178
Text abstractness	Concreteness	Mean for content words	**413.328**	**394.065**	**0.029***
	Mean hypernym	Values of nouns	5.046	4.873	0.307
		Values of verbs	1.648	1.632	0.936

*Note: *p < .05; **p < .01*

Results from expert judgement on contextual features

Based on the 12 judges' responses to the Contextual Parameter Proforma (see Table 10 and Appendix 2), most of the texts in the GEPT reading tests at the B1 and B2 levels were in the social domain; see Figure 6. The two tests appeared to contain texts mostly belonging to the magazine and newspaper article/report genre; see Figure 7. The GEPT Intermediate texts were mostly expository while the GEPT High-Intermediate contained a wider variety of rhetorical organisations; see Figure 8.

Figure 6 Distribution of text domains by test

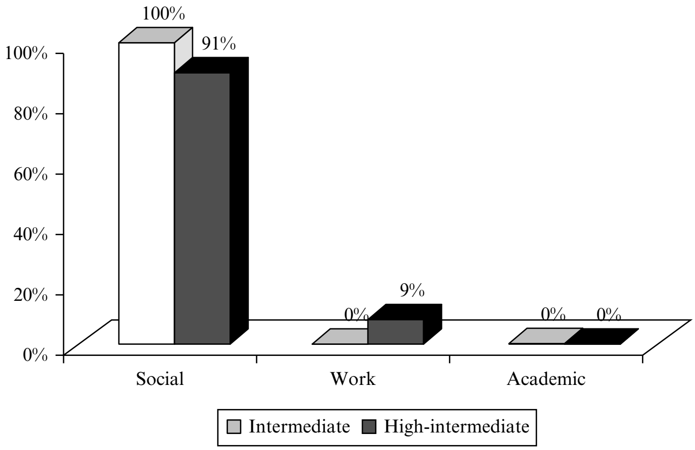
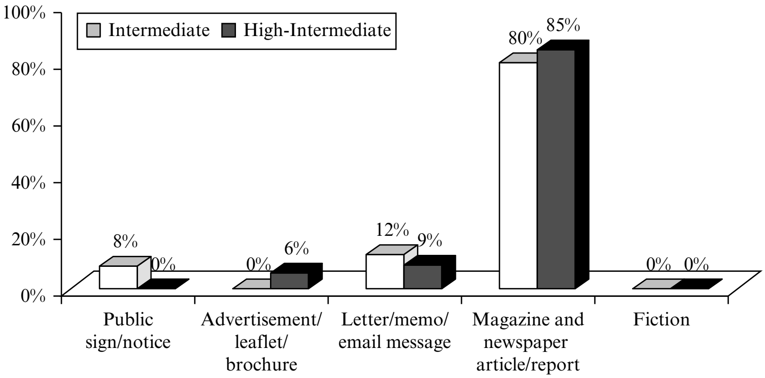


Figure 7 Distribution of text genres by test



In terms of explicitness of rhetorical organisation, text abstractness, subject specificity, and cultural specificity of the GEPT reading texts at the B1 and B2 levels, responses from the expert judgement fell toward the lower end of a 5-point Likert scale (see Figures 9 to 12). However, the judges considered that the GEPT texts at the B1 level were more explicitly organised, semantically more concrete, and more subject and cultural neutral than those at the B2 level; that is to say, the higher the GEPT level, the higher degree of organisation implicitness, text abstractness, subject specificity, and cultural specificity.

Figure 8 Distribution of rhetorical organisations by test

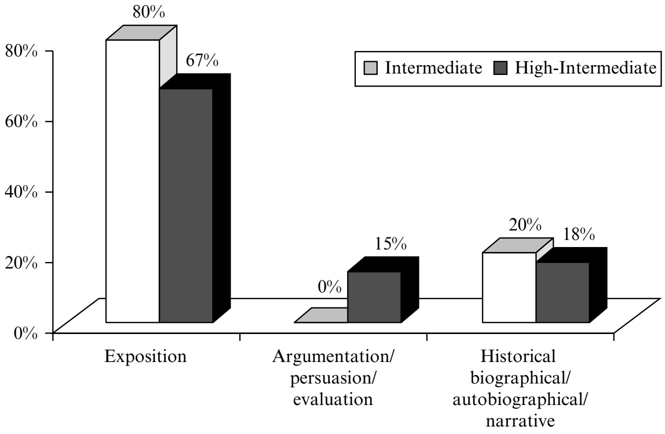


Figure 9 Degree of explicitness of rhetorical organisations by test

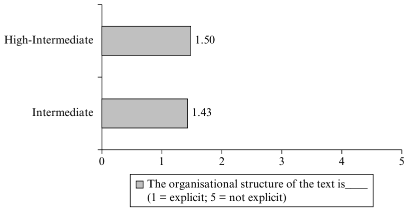


Figure 10 Degree of text abstractness by test

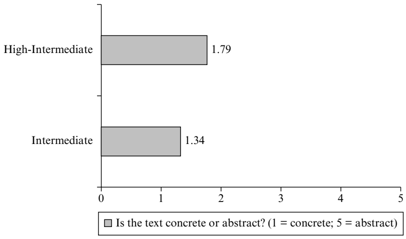


Figure 11 Degree of subject specificity by test

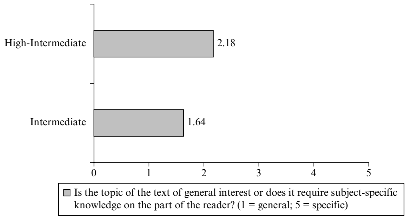
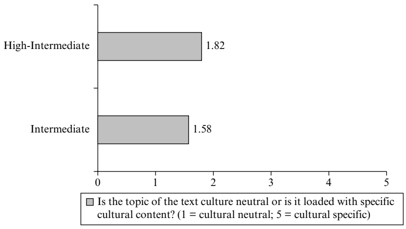


Figure 12 Degree of cultural specificity by test



With respect to the item dimensions, around 80% of the test questions from GEPT reading papers at the B1 and B2 levels were specific detail questions and over 90% of the test questions from the two GEPT tests were factual questions, for which the test takers could find the answer from explicitly stated information in the texts. See Figures 13 and 14.

Figure 13 Type of comprehension questions by test

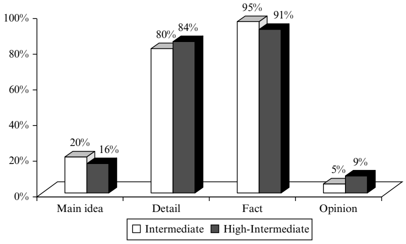
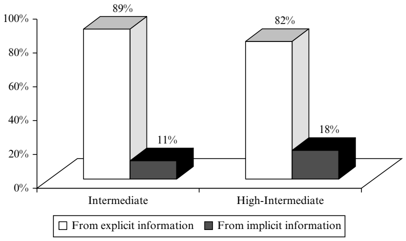
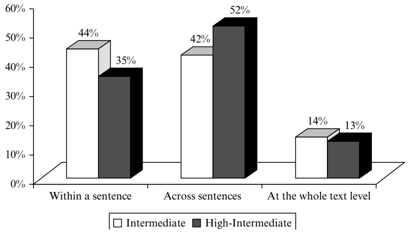


Figure 14 Questions required comprehension of textually explicit or implicit information



The GEPT tests at the B1 level contained a higher proportion of questions that required local comprehension, i.e. within a sentence, than the GEPT tests at the B2 level. Based on expert judgement, the GEPT tests at both the B1 and B2 levels contained fewer than 20% of the items involving comprehension at the whole text level; see Figure 15.

Figure 15 Scope of text content needed to process by test

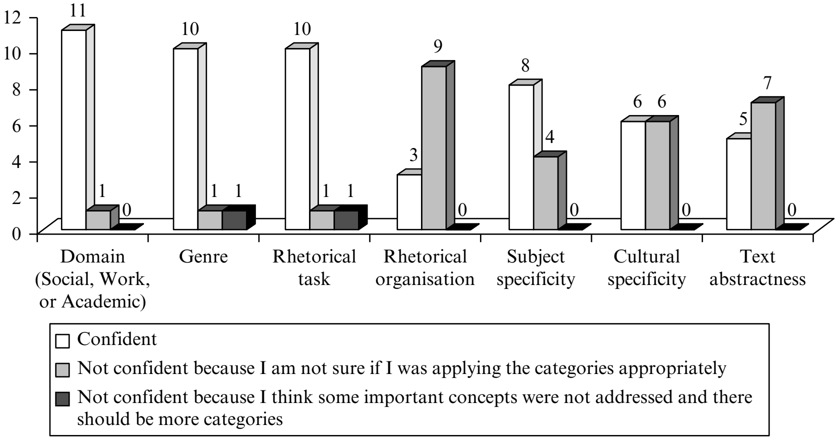


Overall, the GEPT reading tests at the B2 level tended to be cognitively more challenging than those at the B1 level: the GEPT tests at the B2 level had slightly more main idea and opinion questions, more questions that required comprehension of textually implicit information, and more questions that required test takers to comprehend across sentences than those at the B1 level.

Feedback from judges on confidence in responding to Contextual Parameter Proforma

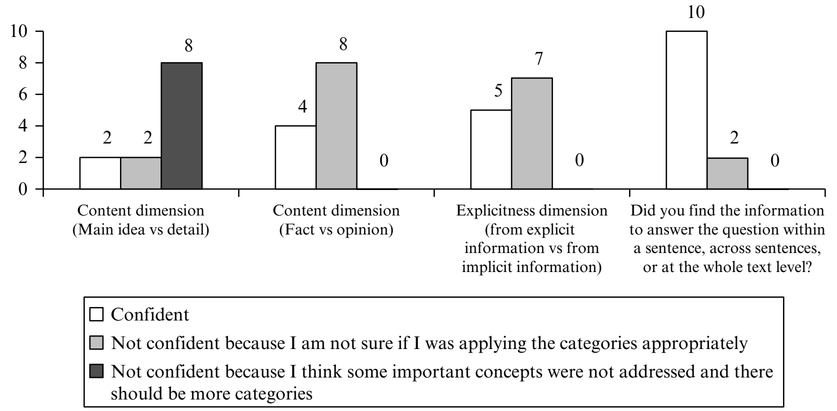
To evaluate how confident the judges felt when making judgements on contextual parameters using the Contextual Parameter Proforma (see Table 10 and Appendix 2), the Feedback Evaluation Questionnaire–2 (see Table 12) was developed. Overall, the participants reported that they were confident when responding to items in the Proforma. However, they were unsure whether they applied the categories appropriately on several text dimension parameters which required judgement on five-point Likert scales, i.e. rhetorical organisation, subject specificity, cultural specificity, and text abstractness; see Figure 16.

Figure 16 Judges' feedback on their confidence in responding to the Contextual Parameter Proforma concerning the text dimension



Concerning the item dimension, they preferred a more common approach to classifying item types, i.e. main idea, detail, inference, and contextual feature, to the dichotomous classification, i.e. main idea vs. detail, and fact vs. opinion, that were in use in the present study; see Figure 17.

Figure 17 Judges' feedback on their confidence in responding to the Contextual Parameter Proforma concerning the item dimension



Feedback from judges on usefulness of the results obtained from the automated textual analysis for determining differences between different test levels

To evaluate how useful the judges felt about the indices obtained from the automated textual analysis for determining differences between adjacent test levels, they responded to the Feedback Evaluation Questionnaire–1; see Table 11. Overall, participants found most of the contextual parameters that were automatically processed, i.e. lexis, syntax, and readability, useful, but indices related to cohesion and text abstractness of limited use; see Figures 18 to 22.

Figure 18 Judges' feedback on usefulness of indices concerning vocabulary

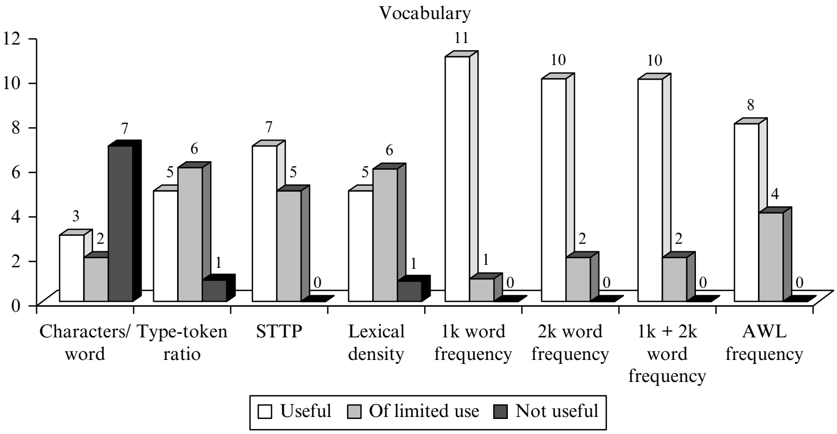


Figure 19 Judges' feedback on usefulness of indices concerning grammar

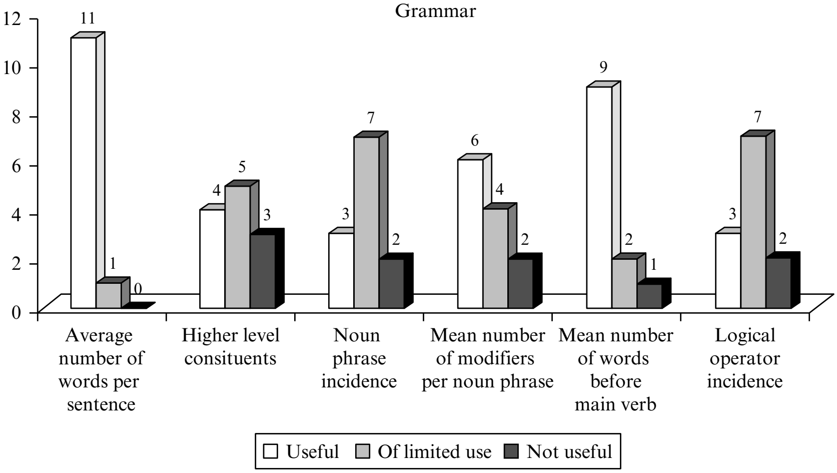


Figure 20 Judges' feedback on usefulness of readability indices

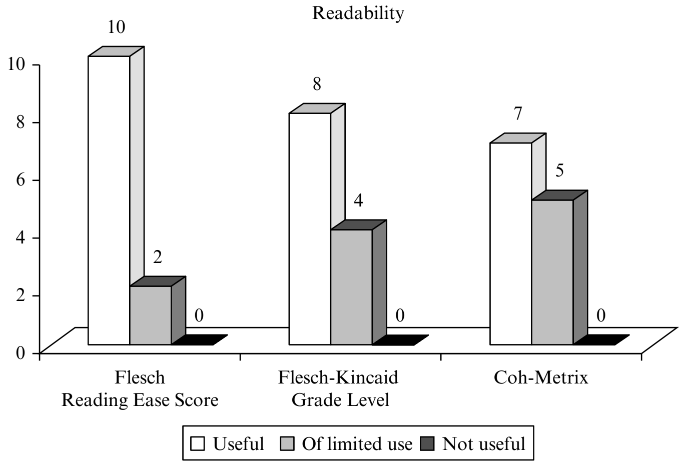


Figure 21 Judges' feedback on usefulness of indices concerning text cohesion

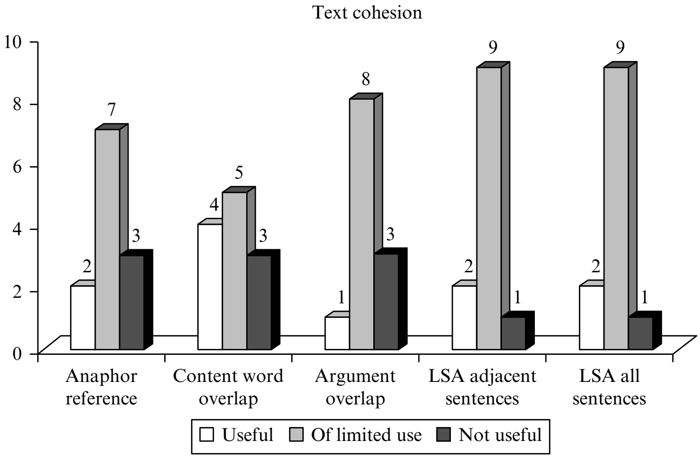
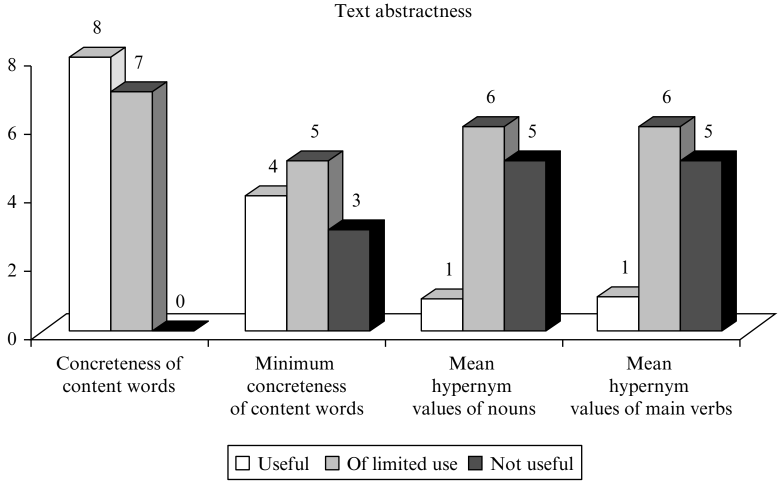


Figure 22 Judges' feedback on usefulness of indices concerning text abstractness



It appeared that the experts tended to judge the tasks based on surface features of the texts intuitively, and they did not find those that needed more sophisticated analysis useful.

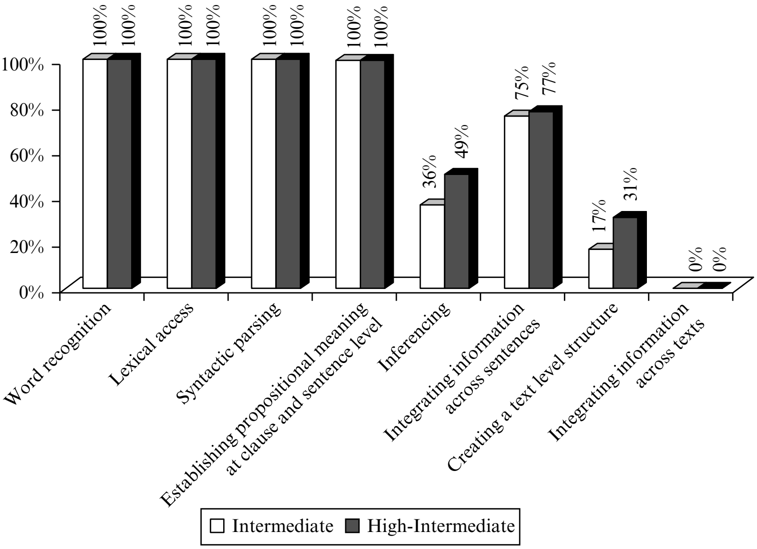
Cognitive processing in reading

Expert judgement on the cognitive processes involved when test takers were taking the GEPT tests at the B1 and B2 levels was quantified using the Cognitive Processing Proforma (see Appendix 3). The responses to the Proforma were weighted based on the tasks' contribution to the total score of the test in question, and then averaged so that tasks with a different number of test questions could be compared on a common basis.

Results from expert judgement on cognitive processes

Based on the judges' responses to the Cognitive Processing Proforma (see Appendix 3), there appeared to be no difference in the four lower order cognitive processing skills (i.e. word recognition, lexical access, syntactic parsing, and establishing propositional meaning at clause and sentence level) between the GEPT reading tests at the B1 and B2 levels. As to the four higher order skills (i.e. inferencing, integrating information across sentences, creating a text level structure, and intergrating information across texts), the judges considered that overall the GEPT reading tests at the B2 level activated test takers to use higher order skills more often than those at the B1 level; see Figure 23.

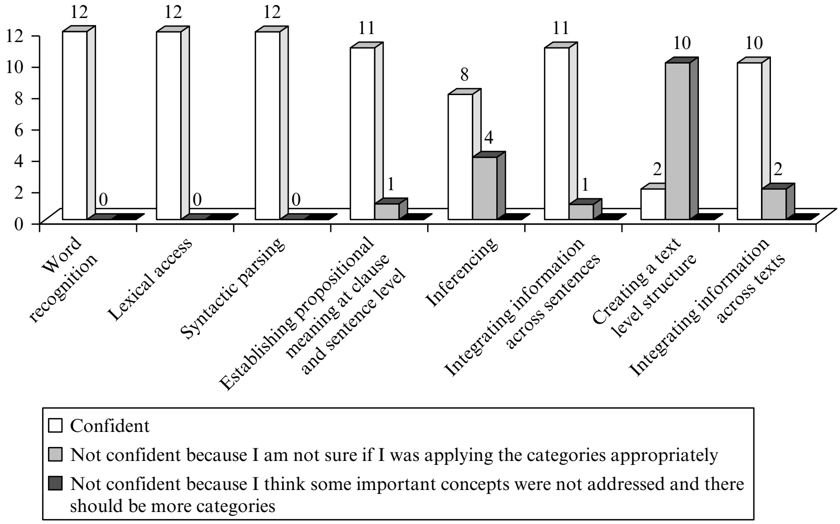
Figure 23 Results from expert judgement on cognitive processes



Feedback from judges on confidence in responding to the Cognitive Processing Proforma

Overall, participants reported that they were confident in determining their responses to most of the items in the Cognitive Processing Proforma (see Appendix 3), but felt unsure whether they had responded to 'creating a text level structure' and also 'inferencing' appropriately; see Figure 24. It is speculated that processes involving inferencing and at the higher discourse construction level occur less frequently in exams at the B1 and B2 level, and it was also likely that the experts were less familiar with these cognitive operations, and therefore they were less disposed to say these occurred.

Figure 24 Results from expert judgement on cognitive processing skills used by test takers



Discussion

To answer Research Question 1 (Is a GEPT reading test designed to measure at CEFR B2 level more difficult than a GEPT reading test designed to measure at CEFR B1 level in terms of test results, contextual parameters, and cognitive processing skills?) both a quantitative approach (i.e. vertically linking scores from different levels of the GEPT onto a common score scale) and a qualitative approach (i.e. contextual parameter and cognitive processing analysis) were adopted. The results from vertical scaling study showed both means of IRT difficulty (*b*) estimates (see Table 21), and those of IRT ability (θ) estimates (see Table 24) increased with the GEPT levels. According to the results from the contextual parameter analysis and the cognitive processing analysis, the GEPT reading tests at the B2 level were lexically more complex, more abstract, and cognitively more challenging than those at the B1 level.

The results from the vertical scaling study and contextual and cognitive processing parameter analyses led to the answer to Research Question 1: a GEPT reading test designed to measure at CEFR B2 level was more difficult than a GEPT reading test designed to measure at CEFR B1 level in terms of test results, contextual parameters, and cognitive processing skills. As the reliability coefficients (Cronbach's α) of GEPT reading tests at different levels were around .83 to .85 (Language Training and Testing Center 2008a:3, 2008b:3, 2008c:4), the GEPT reading papers appeared to generate consistent test results, and these results should be generalisable over occasions.

5 Results and discussion 2: Horizontal comparisons between the GEPT and Cambridge English reading tests at the same CEFR levels

Chapter overview

This chapter reports results on horizontal comparisons between the GEPT and Cambridge English reading tests at the same CEFR level to answer Research Question 2 (Are GEPT reading tests at CEFR B1 and B2 levels comparable to alternative CEFR-linked measures in terms of test results, contextual parameters, and cognitive processing skills?). The extent to which the GEPT is comparable with another CEFR-linked measure, the Cambridge English Reading tests, targeting the same levels, supports the interpretation that the GEPT is appropriate to the level it intends to measure.

Empirical relationships between the scores on the GEPT and Cambridge English reading tests at the two levels are first discussed, followed by the results from both expert judgement on and automated textual analysis of the contextual parameters that affect the comprehensibility and difficulty of reading tasks. Results from analysis of the cognitive operations involved when examinees were taking the GEPT and Cambridge English tests from both experts' and test takers' perspectives are then discussed.

Comparisons between test scores

GEPT and Cambridge English reading tests at CEFR B1 level

A total of 132 target examinees of the GEPT Intermediate level took both the GEPT Intermediate and Cambridge English PET reading tests, both targeting CEFR B1 level, in two consecutive test sessions with a 10-minute break in between. To minimise any practice effect, a single group counter-balanced design was used. The students were randomly divided into four groups: Groups 1 and 2 took reading tests only, and Groups 3 and 4 took reading tests and, in addition, were asked to fill out the Cognitive Processing Checklist (henceforth 'the Checklist'; see Table 17 in Chapter 3 and Appendix 1) immediately after they answered each comprehension question (see Table 8 for data collection design).

To compare test results between the GEPT Intermediate level and Cambridge English PET reading tests on a common basis, the Cambridge English PET was scored based on the number of items correct (the same procedure as the Cambridge English PET scoring scheme) and converted to a 120 point scale, which the GEPT uses operationally.

To examine whether different orders of administering the two tests and whether administering the tests together with the Checklist affected test takers' performance, ANOVA was performed. No significant difference was found among test results of the four groups (see Table 27), suggesting different orders of administering the tests and administering the tests along with the Checklist did not affect test takers' performance on the GEPT Intermediate and Cambridge English PET reading tests. For details, see Tables 26 and 27.

Table 26 Descriptive statistics on the counter-balanced design of the GEPT and Cambridge English reading tests at CEFR B1 level

						95% Confidence Interval for Mean			
		N	Mean	SD	Std error	Lower bound	Upper bound	Min	Max
GEPT Intermediate	**Group 1**	21	71.00	17.085	3.728	63.22	78.78	51	111
	Group 2	40	65.55	17.368	2.746	60.00	71.10	30	108
	Group 3	36	68.50	15.671	2.612	63.20	73.80	45	105
	Group 4	35	71.23	16.835	2.846	65.45	77.01	27	111
Cambridge PET	**Group 1**	21	70.24	17.326	3.781	62.35	78.12	45	103
	Group 2	40	66.58	17.506	2.768	60.98	72.17	24	99
	Group 3	36	65.42	18.080	3.013	59.30	71.53	34	106
	Group 4	35	71.94	18.752	3.170	65.50	78.38	27	110

Table 27 One way ANOVA for the GEPT and Cambridge English tests at the B1 level

		Sum of squares	df	Mean square	F	Sig.
GEPT Intermediate	**Between groups**	733.110	3	244.370	0.873	0.457
	Within group	35833.071	128	279.946		
	Total	36566.182	131			
Cambridge PET	**Between groups**	961.499	3	320.500	0.992	0.399
	Within group	41352.220	128	323.064		
	Total	42313.720	131			

Since no significant difference was observed in test results of the four groups, all scores were pooled together so that the analysis could be performed based on a larger sample. The 132 test takers scored 68.73 on the GEPT Intermediate and 68.27 on the Cambridge English PET. No significant difference (*p*=.69) was observed between test takers' performance on the GEPT Intermediate and Cambridge English PET reading tests. Pearson correlation coefficient of 0.69 indicated a moderate correlation between the GEPT Intermediate and Cambridge English PET reading test scores. For details, see Table 28.

Table 28 Statistics on test takers' performance on the GEPT and Cambridge English tests at the B1 level

	GEPT Intermediate	Cambridge PET
N	132	
Mean	68.73	68.27
Standard deviation	16.71	17.97
% correct	0.57	0.57
Max.	111	110
Min.	27	24
Degree of freedom	(1,131)	
F	0.15	
Critical value (2-tailed)	3.90	
***p* (2-tailed)**	0.69	
Correlation (Pearson)	0.69	

Compared to the test results of the same GEPT Intermediate reading test when it was administered operationally, i.e. 77.95 (N=57,108), the sample test takers scored around 9 points lower than the operational test takers, suggesting that reading proficiency of the examinees was relatively lower than that of the examinees of the operational GEPT Intermediate test.

Based on the test results, the GEPT and Cambridge English reading tests at the B1 level were comparable. The distributions of test takers' scores from the two tests were both symmetric. See Figure 25 for the score distributions and Figure 26 for the scatter plots for scores from the GEPT Intermediate and Cambridge English PET reading tests.

GEPT and Cambridge English reading tests at CEFR B2 level

A total of 138 target examinees of the GEPT High-Intermediate level took both the GEPT High-Intermediate and Cambridge English FCE reading tests, both targeting the B2 level, in two consecutive test sessions with a

Figure 25 Score distributions of the GEPT and Cambridge English reading tests at CEFR B1 level

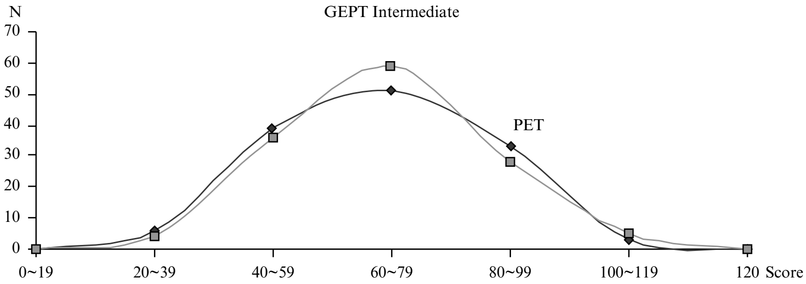
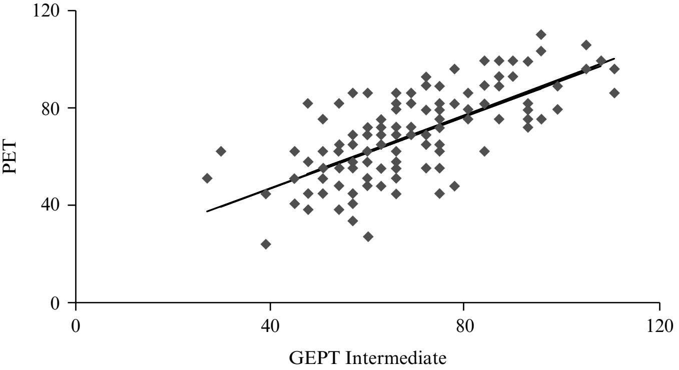


Figure 26 Scatter plots for scores from the GEPT and Cambridge English reading tests at CEFR B1 level



10-minute break in between. To minimise any practice effect, a single-group counter-balanced design was applied. The students were randomly divided into four groups: Groups 1 and 2 took reading tests only, and Groups 3 and 4 took reading tests and were asked to fill out the Checklist immediately after they answered each comprehension question (see Table 8 in Chapter 3 for data collection design).

To compare test results between the GEPT High-Intermediate and Cambridge English FCE reading tests on a common basis, examinees' responses to the FCE Reading test were calculated based on the number of items correct and the marking scheme for the FCE Reading paper (each correct answer in Parts 1 and 2 receives 2 points and each correct answer in Part 3 receives 1 point), and then converted to a 120 point scale which the GEPT uses operationally.

ANOVA was performed to examine whether different orders of administering the two tests and whether administering the tests together with the Checklist affected test takers' performance. No significant difference was found among the test results of the four groups (see Table 30), suggesting the different orders of administering the tests and administering the tests along with the Checklist did not affect test takers' performance on the GEPT High-Intermediate and Cambridge English FCE reading tests. For details, see Tables 29 and 30.

Table 29 Descriptive statistics for the GEPT and Cambridge English reading tests at CEFR B2 level

						95% Confidence Interval for Mean			
		N	Mean	SD	Std error	Lower bound	Upper bound	Min	Max
GEPT High-Intermediate	**Group 1**	31	81.00	14.327	2.573	75.74	86.26	56	117
	Group 2	34	79.09	15.146	2.598	73.80	84.37	43	115
	Group 3	31	78.74	15.669	2.814	72.99	84.49	56	117
	Group 4	42	81.19	12.526	1.933	77.29	85.09	59	104
Cambridge FCE	**Group 1**	31	68.39	21.884	3.930	60.36	76.41	27	109
	Group 2	34	67.68	19.140	3.283	61.00	74.35	35	115
	Group 3	31	66.35	18.748	3.367	59.48	73.23	29	117
	Group 4	42	68.29	15.549	2.399	63.44	73.13	40	101

Table 30 One way ANOVA for the GEPT and Cambridge English reading tests at the B2 level

		Sum of squares	df	Mean square	F	Sig.
GEPT High-Intermediate	**Between groups**	166.976	3	55.659	0.271	0.846
	Within group	27527.147	134	205.426		
	Total	27694.123	137			
Cambridge FCE	**Between groups**	85.072	3	28.357	0.081	0.970
	Within group	46914.464	134	350.108		
	Total	46999.536	137			

Since no significant difference was observed, all test takers' performance was then pooled together for analysis. The 138 test takers scored 80.08 on the GEPT High-Intermediate and 67.72 on the FCE. Significant difference ($p<.05$) was observed between test takers' performance on the GEPT and Cambridge English reading tests at the CEFR B2 level. Pearson correlation coefficient, 0.58, suggested a moderate correlation between the GEPT High-Intermediate and FCE reading test scores. For details, see Table 31.

Table 31 Statistics on test takers' performance on the GEPT and Cambridge English reading tests at the B2 level

	GEPT High-Intermediate	Cambridge FCE
N	138	
Mean	80.08	67.72
Standard deviation	17.95	18.52
% correct	0.67	0.56
Max.	117	101
Min.	43	27
Degree of freedom	(1,137)	
F	38.69	
Critical value (2-tailed)	3.90	
***p* (2-tailed)**	0.00**	
Correlation (Pearson)	0.58	

p<.05

The sample test takers scored a mean of 80.08, very close to the test takers' performance, i.e. 80.02 (N=9,852), on the same GEPT High-Intermediate reading test when it was administered operationally, suggesting that reading proficiency of the sample examinees was very close to the examinees of the operational GEPT High-Intermediate test.

Based on these test results, the Cambridge English FCE was significantly more difficult than the GEPT High-Intermediate reading test. The distribution of test takers' scores on the Cambridge English FCE was symmetric, while that of the GEPT High-Intermediate was negatively skewed. See Figure 27 for the score distributions and Figure 28 for the scatter plots for test takers' performance on the GEPT and Cambridge English reading tests at B2 level.

Figure 27 Score distributions of the GEPT and Cambridge English reading tests at CEFR B2 level

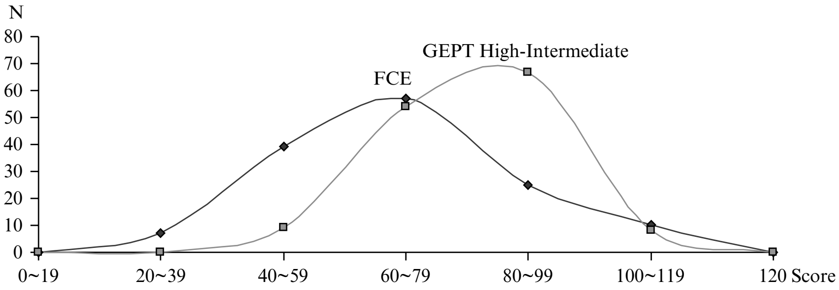
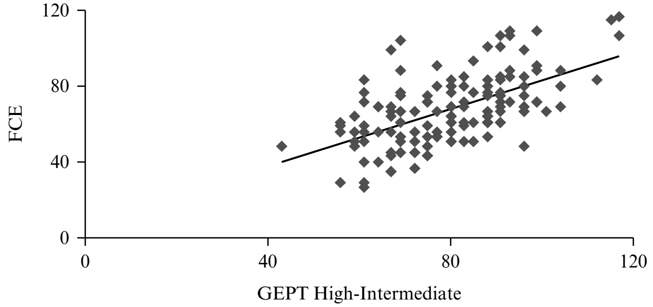


Figure 28 Scatter plots for test scores from the GEPT and Cambridge English reading tests at B2 level



Contextual parameter analysis

Textual features (identified in Table 3 in Chapter 2) that affect the comprehensibility and difficulty of reading tasks were analysed using both automated tools and expert judgement. Automated textual analysis, through Coh-Metrix version 2.1 (Graesser et al 2004, McNamara et al 2002), VocabProfile version 6.2 (Cobb 2010), and WordSmith version 5.0 (Scott 2009), was carried out to analyse the GEPT and Cambridge English texts at the B1 and B2 levels. After the indices of each individual text were obtained, they were averaged and tested by Mann-Whitney U Test, a non-parameter t-test, to determine whether the observed differences reached significance.

As regards the textual characteristics that are not measurable by the automated tools, expert judgement was employed to analyse GEPT and Cambridge English tasks at the B1 and B2 levels, using the Contextual Parameter Proforma (henceforth 'the Proforma'; Appendix 2). The responses to the Proforma were weighted based on the tasks' contribution to the total score of the test in question, and then averaged so that tasks with a different number of test questions could be compared on a common basis.

Automated textual analysis

Various aspects of lexical and syntactic complexity, readability, text cohesion, and text abstractness (see Table 9 in Chapter 3) were analysed using the automated tools. The indices obtained from the analysis were compared for:

1. The GEPT texts at the B1 and B2 levels and the Cambridge English texts at the same two levels (see Table 32).
2. The GEPT and Cambridge English texts at the B1 level (see Tables 33 and 34).
3. The GEPT and Cambridge English texts at the B2 level (see Tables 33 and 35).

These factors were considered to determine whether the four pairs of the tests were significantly different in textual features and to identify criterial features that might be useful in describing different test levels.

To investigate whether the GEPT and Cambridge English texts at the B2 level were more difficult than their B1 level counterparts, the Mann-Whitney U Test, was performed between indices obtained based on the GEPT texts at the B1 and B2 levels and between the Cambridge English tests at the same two levels (see Table 32). Statistically significant differences (p<.05) in text length and text abstractness were observed. Both testing systems used longer and more abstract texts (see Text length and Concreteness, Mean for content words) in the B2 level reading papers than in the B1 level papers. On the other hand, different testing systems appeared to have different rationales of test design in terms of other textual features. The GEPT texts at the B2 level contained longer words (see Character/word), a wider range of lexis (see 1k+2k word frequency, AWL frequency, and Off-List words), and longer sentences (see Average number of words/sentence) than those at the B1 level; furthermore, the B2 texts were more difficult to read, in terms of FREs and FK Grade Level readability, than the B1 texts.

As regards the Cambridge English tests, the B2 texts contained syntactically more complex sentences (see Mean number of higher level constituents and Mean number of words before main verb of main clause) than the B1 texts. Nevertheless, unexpectedly, other indices which reached significance suggested the Cambridge English texts at the B2 level were easier than those at the B1 level. The STTR and lexical density of Cambridge English B2 level texts were lower than those of Cambridge English B1 level texts, suggesting the B2 level texts were lexically less complex than the B1 level texts. Various cohesion indices suggested Cambridge English B2 level texts were easier and more cohesive than Cambridge English B1 level texts. For details see Table 32.

To determine whether the GEPT reading texts at CEFR B1 and B2 levels were comparable to the Cambridge English counterparts, the Mann-Whitney U test was again performed between GEPT and Cambridge English texts at the B1 level and also those at the B2 level (see Table 33).

Overall, Cambridge English texts at both the B1 and B2 levels contained more words appearing among the most frequent 2,000 words (see 1k+2k word frequency), fewer words in the Academic Word List (see AWL frequency), but more abstract content words (see Concreteness, mean for content words) than the GEPT texts did. In terms of syntax, the Cambridge English B1 and B2 texts contained longer but less complex sentences, and used a greater variety of sentence structures (see Average number of words/sentence, Mean number of words before main verb, and Sentence syntax similarity) than the GEPT texts at the same level did. For details, see Table 33.

The GEPT and Cambridge English texts at the B1 level were comparable (see Table 34), in terms of word length (see Characters/word), lexical

Table 32 Results on the comparison between the GEPT and Cambridge English texts at CEFR B1 and B2 levels based on the automated textual analysis

Contextual parameter		GEPT			Cambridge English		
		GEPT-I	GEPT-HI	Sig.	PET	FCE	Sig.
Text length		**148.190**	**222.632**	**0.002***	**113.121**	**346.667**	**0.000***
Lexical complexity	Characters/word	**4.596**	**4.929**	**0.001***	4.528	4.523	0.200
	1k word frequency	**77.93%**	**74.14%**	**0.002***	81.09%	81.81%	0.397
	1k+2k word frequency	**85.78%**	**80.84%**	**0.002***	89.24%	88.07%	0.430
	AWL frequency	**3.72%**	**5.95%**	**0.008***	2.82%	3.12%	0.273
	Off-List words	**10.50%**	**13.21%**	**0.036***	7.91%	8.81%	0.356
	STTR	70.682	72.530	0.258	**78.564**	**71.195**	**0.000***
	Lexical density	0.575	0.571	0.649	**0.563**	**0.502**	**0.000***
Syntactic complexity	Average number of words/sentence	**15.655**	**19.234**	**0.011***	19.085	22.912	0.347
	Noun Phrase Incidence Score (per 1,000 words)	283.404	276.763	0.668	281.449	276.591	0.372
	Mean number of modifiers per noun phrase	0.883	0.928	0.247	**0.964**	**0.777**	**0.010***
	Mean number of higher level constituents per sentence	0.724	0.715	0.361	**0.698**	**0.753**	**0.001***
	Mean number of words before main verb of main clause in sentences	4.585	4.821	0.789	**2.555**	**3.480**	**0.032***
	Sentence syntax similarity, all, across paragraphs	0.110	0.112	0.872	0.068	0.078	0.235
Readability	Flesch Reading Ease Score	**64.126**	**49.656**	**0.000***	66.128	65.027	0.834

Table 32 (continued)

Contextual parameter			**GEPT**			**Cambridge English**		
		GEPT-I	**GEPT-HI**	**Sig.**	**PET**	**FCE**	**Sig.**	
Readability	Flesch-Kincaid Grade Level		**8.114**	**10.471**	**0.000***	7.481	8.523	0.119
	Coh-Metrix readability		14.200	13.133	0.320	**12.665**	**17.716**	**0.004***
Cohesion	Referential cohesion	Ratio of pronouns to noun phrases	0.204	0.146	0.145	**0.202**	**0.343**	**0.001***
		Anaphor reference, all distances, unweighted	0.178	0.108	0.347	**0.149**	**0.290**	**0.003***
		Argument overlap, all distances, unweighted	0.449	0.503	0.555	**0.321**	**0.498**	**0.010***
		Stem overlap, all distances, unweighted	0.421	0.521	0.205	0.306	0.354	0.732
		Proportion of content words that overlap between adjacent sentences	0.075	0.078	0.728	**0.066**	**0.100**	**0.029***
	Conceptual cohesion	Logical operator incidence score (and + if + or + cond + neg)	33.624	35.080	0.452	39.536	41.385	0.537
		LSA adjacent sentences	0.212	0.246	0.178	0.168	0.148	0.419
		LSA all sentences	0.227	0.261	0.178	0.168	0.133	0.274
Text abstractness	Concreteness	Mean for content words	**413.328**	**394.065**	**0.029***	**399.100**	**374.772**	**0.007***
	Mean hypernym	Values of nouns	5.046	4.873	0.307	**5.158**	**4.819**	**0.030***
		Values of verbs	1.648	1.632	0.936	1.515	1.617	0.091

Note: **p*<.05; ***p*<.01.

Table 33 Results on the comparison between the GEPT and Cambridge English tests at the same CEFR level based on the automated textual analysis

Contextual parameter		B1			B2		
		GEPT-I	Cambridge PET	Sig.	GEPT-HI	FCE	Sig.
Text length		**148.190**	**113.121**	**0.001***	222.632	346.667	0.525
Lexical complexity	Characters/word	4.596	4.528	0.389	**4.929**	**4.523**	**0.000***
	1K word frequency	77.93%	81.09%	0.064	**74.14%**	**81.81%**	**0.000***
	1k+2k word frequency	**85.78%**	**89.24%**	**0.015***	**80.84%**	**88.07%**	**0.000***
	AWL frequency	3.72%	2.82%	0.187	**5.95%**	**3.12%**	**0.001***
	Off-List words	**10.50%**	**7.91%**	**0.030***	**13.21%**	**8.81%**	**0.000***
	STTR	**70.682**	**78.564**	**0.000***	72.530	71.195	0.402
	Lexical density	0.575	0.563	0.270	**0.571**	**0.502**	**0.000***
Syntactic complexity	Average number of words/sentence	**15.655**	**19.085**	**0.044***	19.234	22.912	0.817
	Noun Phrase Incidence Score (per 1,000 words)	283.404	281.449	0.404	276.763	276.591	0.863
	Mean number of modifiers per noun phrase	0.883	0.964	0.851	**0.928**	**0.777**	**0.009***
	Mean number of higher level constituents per sentence	0.724	0.698	0.213	**0.715**	**0.753**	**0.004***
	Mean number of words before main verb of main clause in sentences	**4.585**	**2.555**	**0.006***	**4.821**	**3.480**	**0.014***
	Sentence syntax similarity, all, across paragraphs	**0.110**	**0.068**	**0.000***	**0.112**	**0.078**	**0.006***
Concreteness	Mean for content words	**413.328**	**399.100**	**0.017***	**394.065**	**374.772**	**0.014***

Note: **p<.05; **p<.01.*

complexity (see AWL frequency and lexical density), syntactic complexity (see Noun Phrase Incidence Score, Mean number of modifiers per noun phrase, and Mean number of higher level constituents), and all readability and cohesion indices. However, a few indices (see STTR, Sentence length, Sentence syntax similarity, and Concreteness, mean for content words) suggested the GEPT B1 level texts were easier than the Cambridge English B1 level tests, while some (see Text length, 1k+2k word frequency, and Mean number of words before main verb of main clause) suggested the GEPT B1 level texts were more difficult than the Cambridge English B1 level texts.

As regards the B2 level tests (see Table 35), most lexical and syntactic complexity indices and all three readability indices (see FRE, FK Grace Level, and Coh-Metrix readability) showed that the GEPT texts were more challenging than the Cambridge English texts. The Cambridge English texts at the B2 level were lexically less complex (see Characters/word, 1k+2k word frequency, AWL frequency, and Lexical density), contained syntactically less complex structures (see Mean number of modifiers per noun phrase and Mean number of words before main verb) and were more cohesive (see Anaphor reference).

On the other hand, a few of the text cohesion (e.g. Ratio of pronouns to noun phrases, Stem overlap, LSA) and abstractness indices (e.g. Concreteness, mean for content words) suggested the Cambridge English texts were more difficult than the GEPT texts: the Cambridge English texts at the B2 level had higher density of pronouns (see Ratio of pronouns to noun phrases) and were conceptually less similar across the text (see LSA indices) than the GEPT counterpart, which were expected to make reading the Cambridge English texts at the B2 level more difficult: a high density of pronouns may cause referential cohesion problems when a reader does not know what the pronouns refer to, while a lower LSA index suggests that the text is less cohesive conceptually and therefore the ease and speed of text processing may be impeded. For details, see Table 35.

To further investigate the differences found between the GEPT and Cambridge English reading texts at CEFR B1 and B2 levels, the indices obtained from Coh-Metrix automated textual analysis were compared with Grade 12 and College Level norm values, respectively, which were provided by Coh-Metrix. The Coh-Metrix norm values were computed based on sample texts from TASA (Touchstone Applied Science Associates) corpus (Landauer, Foltz and Laham 1998), consisting of over 10 million words from random samples of texts that students in the USA read.

Significant differences ($p<.05$) were found between the GEPT and Cambridge English texts at the same CEFR level (see shaded areas in Table 36). Most indices showed that the features of the GEPT texts at both the B1 and B2 levels were closer to those of the texts in Coh-Metrix Grade 12 and College level norms, respectively, in terms of grammatical structures, readability,

Table 34 Results on the comparison between the GEPT and Cambridge English tests at CEFR B1 level based on the automated textual analysis

Contextual parameter		B1				
		GEPT-I		Cambridge PET		Sig.
Text length		**148.190**	+	**113.121**	−	**0.001***
Lexical complexity	Characters/word	4.596		4.528		0.389
	1k word frequency	77.93%		81.09%		0.064
	1k+2k word frequency	**85.78%**	+	**89.24%**	−	**0.015***
	AWL frequency	3.72%		2.82%		0.187
	STTR	**70.682**	−	**78.564**	+	**0.000***
	Lexical density	0.575		0.563		0.270
Syntactic complexity	Average number of words/sentence	**15.655**	−	**19.085**	+	**0.044***
	Noun Phrase Incidence Score (per 1,000 words)	283.404		281.449		0.404
	Mean number of modifiers per noun phrase	0.883		0.964		0.851
	Mean number of higher level constituents per sentence	0.724		0.698		0.213
	Mean number of words before main verb of main clause in sentences	**4.585**	+	**2.555**	−	**0.006***
	Sentence syntax similarity, all, across paragraphs	**0.110**	−	**0.068**	+	**0.000***

Table 34 (continued)

Readability	Flesch Reading Ease Score		64.126		66.128		0.534
	Flesch-Kincaid Grade Level		8.114		7.481		0.459
	Coh-Metrix readability		14.200		12.665		0.534
Cohesion	Referential cohesion	Ratio of pronouns to noun phrases	0.204		0.202		0.981
		Anaphor reference, all distances, unweighted	0.178		0.149		0.740
		Argument overlap, all distances, unweighted	0.449		0.321		0.075
		Stem overlap, all distances, unweighted	0.421		0.306		0.097
		Proportion of content words that overlap between adjacent sentences	0.075		0.066		0.404
	Conceptual cohesion	Logical operator incidence score (and + if + or + cond + neg)	33.624		39.536		0.672
		LSA adjacent sentences	0.212		0.168		0.335
		LSA all sentences	0.227		0.168		0.235
Concreteness hypernym	Mean for content words		**413.328**	**−**	**399.100**	+	**0.017***

Note: − stands for easier, + stands for more difficult
**p<.05; **p<.01.*

Table 35 Results on the comparison between the GEPT and Cambridge English tests at CEFR B2 levels based on the automated textual analysis

Contextual parameter		B2				
		GEPT-HI		FCE		Sig.
Text length		222.632		346.667		0.525
Lexical complexity	Characters/word	**4.929**	+	**4.523**	−	**0.000***
	1k word frequency	**74.14%**	+	**81.81%**	−	**0.000***
	1k+2k word frequency	**80.84%**	+	**88.07%**	−	**0.000***
	AWL frequency	**5.95%**	+	**3.12%**	−	**0.001***
	STTR	72.530		71.195		0.402
	Lexical density	**0.571**	+	**0.502**	−	**0.000***
Syntactic complexity	Average number of words/sentence	19.234		22.912		0.817
	Noun Phrase Incidence Score (per 1,000 words)	276.763		276.591		0.863
	Mean number of modifiers per noun phrase	**0.928**	+	**0.777**	−	**0.009***
	Mean number of higher level constituents	**0.715**	−	**0.753**	+	**0.004***
	Mean number of words before main verb of main clause in sentences	**4.821**	+	**3.480**	−	**0.014***
	Sentence syntax similarity, all, across paragraphs	**0.112**	−	**0.078**	+	**0.006***

Table 35 (continued)

Readability	Flesch Reading Ease Score		**49.656**	+	**65.027**	−	**0.001***
	Flesch-Kincaid Grade Level		**10.471**	+	**8.523**	−	**0.006***
	Coh-Metrix readability		**13.133**	+	**17.716**	−	**0.013***
Cohesion	Referential cohesion	Ratio of pronouns to noun phrases	**0.146**	−	**0.343**	+	**0.000***
		Anaphor reference, all distances, unweighted	**0.108**	+	**0.290**	−	**0.000***
		Argument overlap, all distances, unweighted	0.503		0.498		0.488
		Stem overlap, all distances, unweighted	**0.521**	−	**0.354**	+	**0.040***
		Proportion of content words that overlap between adjacent sentences	0.078		0.100		0.080
	Conceptual cohesion	Logical operator incidence score (and + if + or + cond + neg)	35.080		41.385		0.172
		LSA adjacent sentences	**0.246**	−	**0.148**	+	**0.000***
		LSA all sentences	**0.261**	−	**0.133**	+	**0.000***
Concreteness	Mean for content words		**394.065**	−	**374.772**	+	**0.014***

Note: −*stands for easier;* +*stands for more difficult*
**p*<.05; ***p*<.01.

Table 36 Results on the comparison between Coh-Metrix indices of GEPT and Cambridge English tests at CEFR B1 and B2 levels and Coh-Metrix norm values

Contextual parameter		B1			Grade 12 Norm	B2			College level norm
		GEPT-I	Cambridge PET	Sig.		GEPT-HI	FCE	Sig.	
Text length		148.190	113.121	**0.001***	299.44	222.632	346.667	0.525	307.07
Syntactic complexity	Average number of words/sentence	15.655	19.085	**0.044***	19.63	19.234	22.912	0.817	22.03
	Noun Phrase Incidence Score (per thousand words)	283.404	281.449	0.404	274.71	276.763	276.591	0.863	270.23
	Mean number of modifiers per noun phrase	0.883	0.964	0.851	0.927	0.928	0.777	**0.009***	0.974
	Mean number of higher level constituents per word	0.724	0.698	0.213	0.718	0.715	0.753	**0.004***	0.707
	Mean number of words before main verb of main clause in sentences	4.585	2.555	**0.006***	5.196	4.821	3.480	**0.014***	6.169
	Logical operator incidence score (and + if + or + cond + neg)	33.624	39.536	0.672	41.736	35.080	41.385	0.172	47.330
	Sentence syntax similarity, all, across paragraphs	0.110	0.068	**0.000***	0.10	0.112	0.078	**0.006***	0.09

Table 36 (continued)

Readability	Flesch Reading Ease Score	64.126	66.128	0.534	52.47	49.656	65.027	**0.001***	40.14
	Flesch-Kincaid Grade Level	8.114	7.481	0.459	10.54	10.471	8.523	**0.006***	11.61
Cohension	Anaphor reference, all distances, unweighted	0.178	0.149	0.740	0.15	0.108	0.290	**0.000***	0.09
	Argument overlap, all distances, unweighted	0.449	0.321	0.075	0.41	0.503	0.498	0.488	0.40
	Stem overlap, all distances, unweighted	0.421	0.306	0.097	0.38	0.521	0.354	**0.040***	0.44
	Proportion of content words that overlap between adjacent sentences	0.075	0.066	0.404	0.10	0.078	0.100	0.080	0.08
	LSA adjacent sentences	0.212	0.168	0.335	0.36	0.246	0.148	**0.000***	0.41
	LSA all sentences	0.227	0.168	0.235	0.31	0.261	0.133	**0.000***	0.36
Text abstractness	Concreteness, mean for content words	413.328	399.100	**0.017***	377.52	394.065	374.772	**0.014***	372.54
	Mean hypernym values of nouns	5.046	5.158	0.169	4.98	4.873	4.819	1.000	4.88
	Mean hypernym values of verbs	1.648	1.515	0.296	1.47	1.632	1.617	0.665	1.48

Note: The indices in the shaded cells are closer to the norm values.
**p<.05; **p<.01.*

and cohesion, than the Cambridge English texts. However, surface level features of Cambridge English texts at the B1 and B2 levels, such as number of connectives (see Logical operator incidence score) and concreteness (see Concreteness, Mean for content words), were closer to the norm than those of the GEPT texts. As regards the two B1 level texts, text length and syntactic complexity indices (see Mean number of words before main verb and Sentence syntax similarity) suggested that GEPT texts were closer to the norm, while concreteness indices showed that Cambridge English PET texts were closer to the norm. As to B2 level texts, except one syntactic complexity index (i.e. Sentence syntax similarity) and one text abstractness index (i.e. Concreteness, Mean for content words), almost all syntactic complexity, readability, and cohesion indices suggested that GEPT texts were closer to the norm.

Expert judgement on contextual features

Based on the 12 judges' responses to Contextual Parameter Proforma (see Appendix 2), most of the texts in the four tests were in the social domain (see Figure 29). The Cambridge English FCE texts contained texts mostly from fiction books, and GEPT texts at both the B1 and B2 levels contained texts mostly belonging to the magazine and newspaper article/report genre (see Figure 30). All the FCE texts were narrative, while the texts from the other three tests were mostly expository and the GEPT High-Intermediate was the only test which had argumentative texts (see Figure 31).

In terms of the explicitness of rhetorical organisation, text abstractness, subject specificity, and cultural specificity of the GEPT and Cambridge English reading papers at the B1 and B2 levels, responses from the expert judgement fell toward the lower end of a 5-point Likert scale; see Figures 32 to 35. Overall, the judges considered that both GEPT and Cambridge English texts at the B1 level were more explicitly organised, semantically

Figure 29 Distribution of text domains by test

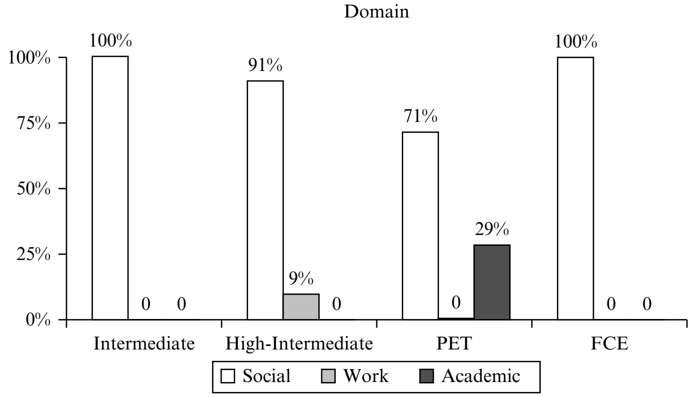


Figure 30 Distribution of text genres by test

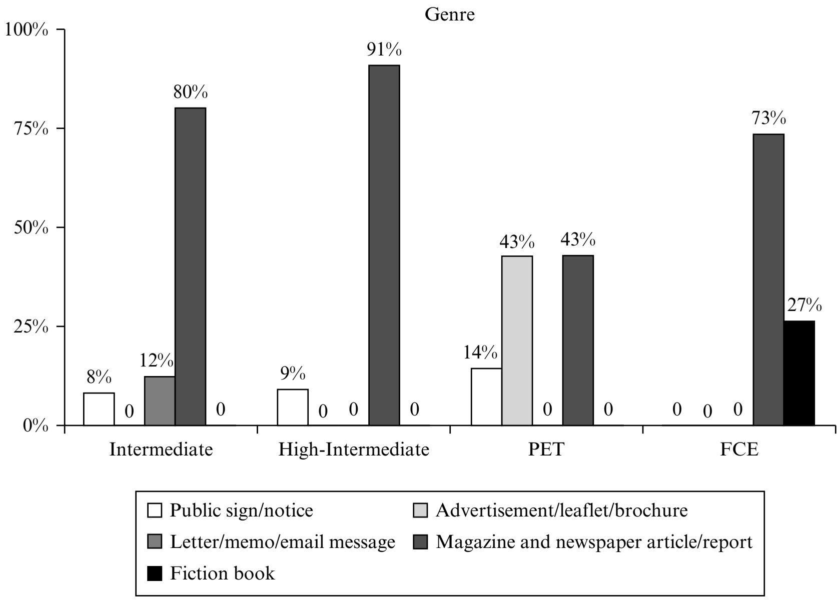
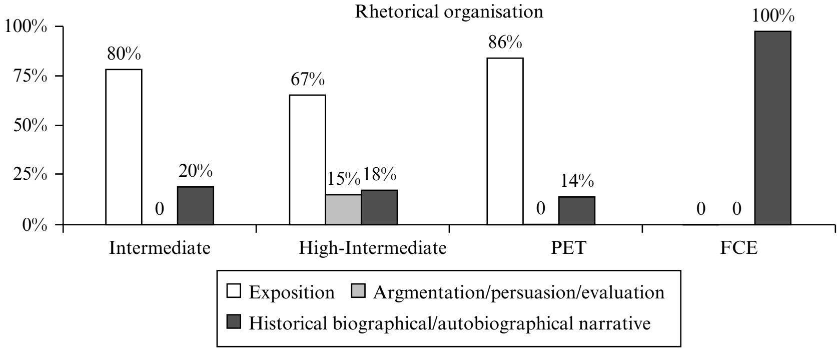


Figure 31 Distribution of rhetorical organisations by test



more concrete, and more subject and cultural neutral than those at B2 level; that is to say, the higher the GEPT and Cambridge English level, the higher degree of organisation implicitness, text abstractness, subject specificity, and cultural specificity.

With respect to the item dimensions, around 80% of the test questions from GEPT and Cambridge English reading papers at the B1 and B2 levels were

Figure 32 Degree of explicitness of rhetorical organisations by test

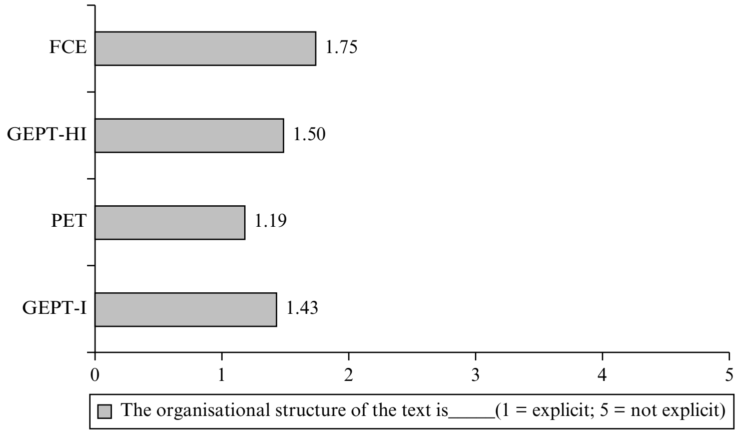
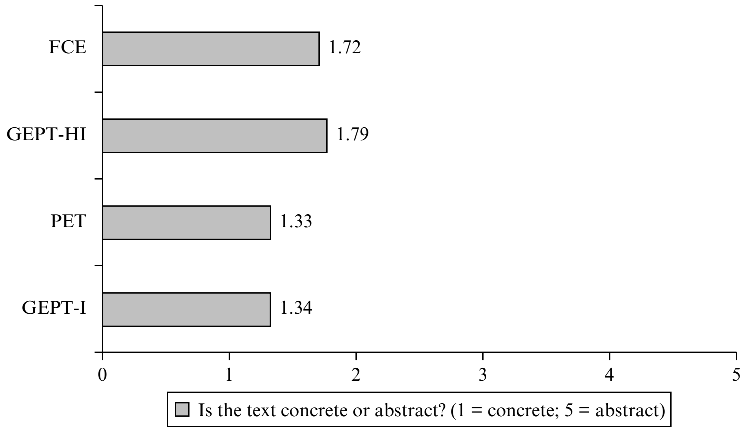


Figure 33 Degree of text abstractness by test



specific detail questions; see Figure 36. The test takers could find the answers to more than 80% of the GEPT Intermediate, the GEPT High-Intermediate, and the Cambridge English PET test questions based on explicitly stated information in the texts, while the Cambridge English FCE contained the most questions that required comprehension of textually implicit information; see Figure 37. Over 90% of the test questions from the GEPT Intermediate, the GEPT High-Intermediate, and the Cambridge English PET tests were factual questions, while the Cambridge English FCE contained the highest proportion of opinion questions; see Figure 38. Both the GEPT and Cambridge

Figure 34 Degree of subject specificity by test

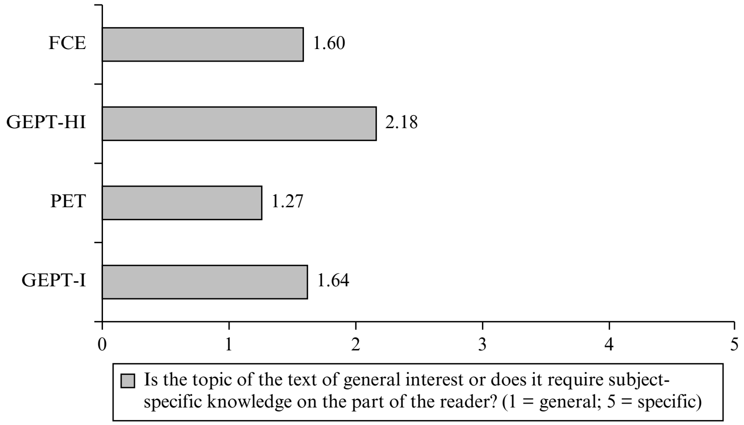
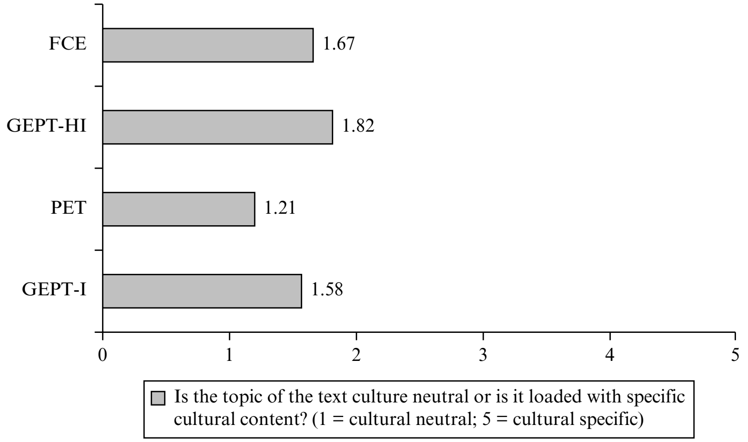


Figure 35 Degree of cultural specificity by test



English tests at the B1 level contained a higher proportion of questions that required local comprehension, i.e. within a sentence, and fewer than 20% of the items involved comprehension at the whole text level; see Figure 39.

Overall, both the GEPT and Cambridge English reading tests at the B2 level tended to have more main idea and opinion questions, more questions requiring comprehension of textually implicit information, and more questions that required test takers to comprehend across sentences or at the whole text level than the B1 level.

Figure 36 Type of comprehension questions by test: Main idea vs detail

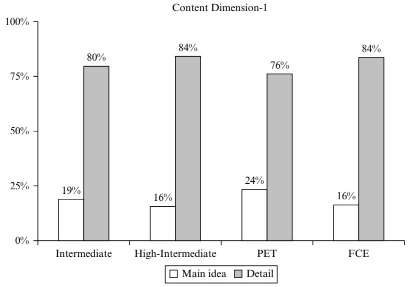


Figure 37 Questions required comprehension of textually explicit or implicit information

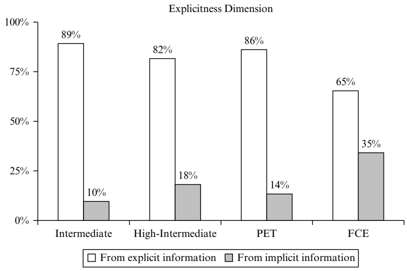


Figure 38 Type of comprehension questions by test: Fact vs opinion

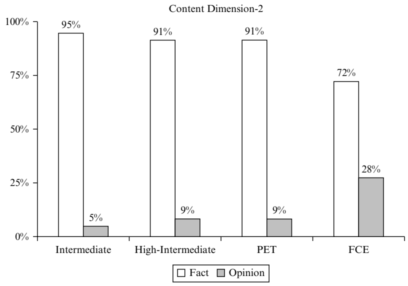
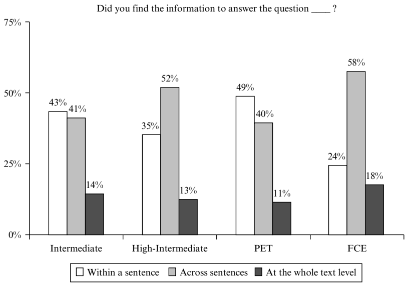


Figure 39 Scope of text content needed to process by test



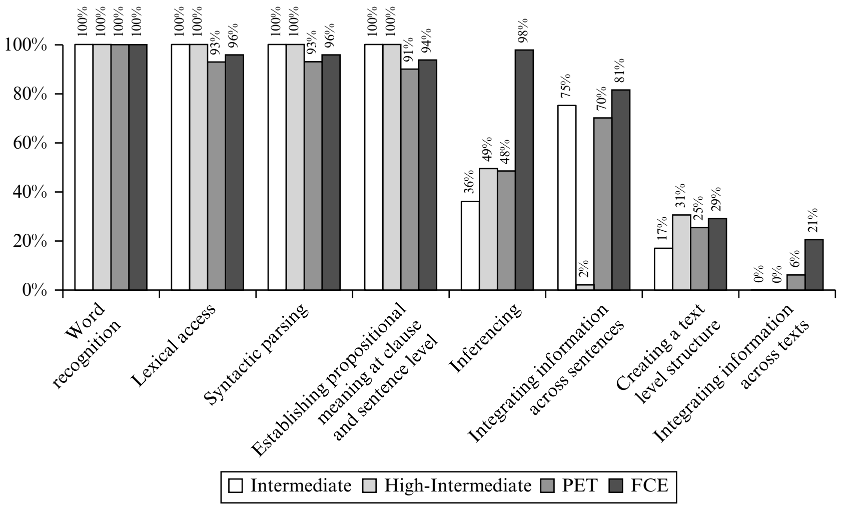
Cognitive processing analysis

Cognitive operations in reading were investigated from both expert judges' and test takers' perspectives. The Cognitive Processing Proforma (henceforth 'the Proforma'; see Table 13 in Chapter 3 and Appendix 3) was employed to quantify expert judgement on the cognitive processes involved when test takers were taking the GEPT and Cambridge English tests at the B1 and B2 levels. The responses to the Proforma were weighted based on the tasks' contribution to the total score of the test in question, and then averaged so that tasks with a different number of test questions could be compared on a common basis. The Cognitive Processing Checklist (see Table 17 in Chapter 3 and Appendix 1) was applied for test takers to report what they actually did to find the answers to each test question. Results from expert judgement and test takers' self-reports were triangulated and are discussed later in this chapter.

Expert judgement on cognitive processes

Based on the expert judges' responses to the Proforma (see Table 13 and Appendix 3), there appeared to be no difference in the four lower order cognitive processing skills (i.e. word recognition, lexical access, syntactic parsing, and establishing propositional meaning at clause and sentence level) between the GEPT and Cambridge English reading tests at the B1 and B2 levels. As to the four higher order skills (i.e. inferencing, integrating information across sentences, creating a text level structure, and intergrating information across texts), the judges considered that overall both the GEPT and Cambridge English reading tests at the B2 level activated test takers to use higher order skills more often than those at the B1 level. In particular, the Cambridge English FCE stimulated test takers to use a higher order skill, i.e. 'inferencing', considerably more often than the other three reading tests. For details, see Figure 40.

Figure 40 Results from expert judgement on cognitive processes



Results from test takers' self-reports on cognitive processes

GEPT and Cambridge English reading tests at B1 level

To investigate what cognitive processing skills test takers used when they were taking the tests, the 71 examinees (Groups 3 and 4 in Table 8 in Chapter 3) who took both the GEPT Intermediate and Cambridge English PET reading tests and filled out the Cognitive Processing Checklist (see Table 17 in

Chapter 3 and Appendix 1) during the tests were first rank-ordered based on their scores of the GEPT Intermediate and Cambridge English PET reading tests, and the highest and the lowest 27% were identified as the High Group (those with high English reading ability) and the Low Group (those with low English reading ability), respectively. The examinees identified as the High Group (based on both GEPT Intermediate and Cambridge English PET scores) and those identified as the Low Group (based on both test scores) were selected for analysis.

A total of 23 out of 71 examinees were selected. The mean scores of the 11 examinees in the High Group were 94.91 and 93.82 on the GEPT Intermediate and on the Cambridge English PET reading test, respectively; and the mean scores of the 12 examinees in the Low Group were 52.00 and 46.66 on the GEPT Intermediate and on the Cambridge English PET reading test, respectively. The t-test result showed significant difference (p=.00) in reading performance on the GEPT Intermediate and on the Cambridge English PET reading test between the High Group and the Low Group. See Table 37.

Table 37 T-test statistics on the High Group and the Low Group of the two CEFR B1 level tests

		High Group (N=11)	Low Group (N=12)	t (*p*)
GEPT Intermediate	Mean	94.91	52.00	11.30
	SD	9.42	8.80	(0.000*)
Cambridge PET	Mean	93.82	46.66	13.67
	SD	8.93	7.97	(0.000*)

**p<.05*

SD=Standard deviation

The mean frequencies, proportions, and standard deviations of each choice of the two items in the Cognitive Processing Checklist (see Table 17 in Chapter 3 and Appendix 1) were computed for the High Group, the Low Group, and the whole group. Overall, examinees applied careful reading (see shaded cells in Table 38) most frequently, except that the High Group reported employing search reading most often when taking the GEPT Intermediate test. The test takers reported that both the GEPT Intermediate and Cambridge PET tasks required text-level comprehension most frequently. For details, see Table 38.

To examine whether differences in the frequencies among the choices of each item within each group reached significance, the Friedman test (a repeated-measures ANOVA) was performed. Significant difference (p<.05) in the frequencies among the five operations under cognitive operations (see

Table 38 Statistics on the mean frequencies, proportions, and standard deviations of each choice of the two items – the reading tests at the CEFR B1 level*

Item			High Group		Low Group		Whole Group	
			Mean (%)	SD	Mean (%)	SD	Mean (%)	SD
GEPT Intermediate	Cognitive processing	A. Scanning	7.45 (19%)	10.31	7.25 (18%)	7.70	9.04 (23%)	9.38
		B. Search reading	16.27 (41%)	13.48	8.50 (21%)	9.59	10.24 (26%)	9.02
		C. From explicitly stated info	6.73 (17%)	10.37	10.25 (26%)	7.45	9.46 (24%)	7.57
		D. From implicit info	7.09 (18%)	9.97	8.08 (20%)	10.13	8.80 (22%)	9.10
		E. Careful reading	12.45 (31%)	8.94	21.75 (54%)	10.25	18.89 (47%)	10.30
	Level of comprehension	A. Intra-sentential	13.91 (35%)	12.62	6.08 (15%)	4.32	10.49 (26%)	9.12
		B. Inter-sentential	10.00 (25%)	6.66	10.00 (25%)	9.11	10.06 (25%)	6.55
		C. Text-level	16.36 (41%)	12.23	24.33 (61%)	9.79	19.80 (50%)	11.09
Cambridge English PET	Cognitive processing	A. Scanning	8.82 (25%)	6.13	6.33 (18%)	6.39	8.55 (24%)	7.61
		B. Search reading	11.82 (34%)	7.15	7.67 (22%)	8.37	10.75 (31%)	6.83
		C. From explicitly stated info	6.00 (17%)	3.19	7.58 (22%)	6.11	7.14 (20%)	6.52
		D. From implicit info	4.36 (12%)	3.47	9.17 (26%)	9.76	7.04 (20%)	6.60
		E. Careful reading	13.82 (39%)	7.37	18.17 (52%)	8.82	15.72 (45%)	8.65
	Level of comprehension	A. Intra-sentential	8.36 (24%)	5.92	4.83 (14%)	4.67	7.96 (23%)	5.88
		B. Inter-sentential	9.45 (27%)	4.97	10.58 (30%)	5.02	10.06 (29%)	5.30
		C. Text-level	17.55 (50%)	8.19	20.25 (58%)	7.37	17.69 (51%)	7.57

**The highest frequency choice within the group are marked in shaded cells.*

Item 1 in Table 17 in Chapter 3 and also in Appendix 1) was observed for the Whole Group and the Low Group (the alpha level of 0.06 of the GEPT Intermediate was very close to the significance level), but not for the High Group. As regards the three operations under Level of comprehension (Item 2 in Table 17 in Chapter 3 and also in Appendix 1), within-group difference reached significance ($p<.05$) for the Whole Group and Low Group when they took the GEPT Intermediate and the Cambridge English PET. See shaded cells in Table 39.

Table 39 Friedman test on the frequencies of the cognitive skills used by different proficiency groups — the CEFR B1 level tests

	Item	Group	df	N	X^2	*p*
GEPT Intermediate	Cognitive processing	Whole	4	71	47.37	.000*
		High	4	11	8.36	.079
		Low	4	12	14.42	.006
	Level of comprehension	Whole	2	71	15.53	.000*
		High	2	11	0.18	.913
		Low	2	12	13.83	.001*
Cambridge PET	Cognitive processing	Whole	4	71	48.21	.000*
		High	4	11	13.05	.011*
		Low	4	12	9.52	.049*
	Level of comprehension	Whole	2	71	32.23	.000*
		High	2	11	2.93	.231
		Low	2	12	15.70	.000*

**p<.05*

The Wilcoxon signed-ranks test was then performed to test within-group differences in frequencies that reached significance based on the Friedman test. The choices were first rank-ordered, and then differences in the frequencies of each pair of consecutive categories were tested. According to the Dunn-Bonferroni correction, the alpha level was adjusted to 0.0125 for Item 1 (see Table 17 in Chapter 3 and also Appendix 1) since the comparison was carried out four times, and the alpha level was adjusted to 0.025 for Item 2 since the comparison was carried out twice.

According to the adjusted alpha level, the results of the Low Group and the Whole Group reached significance. The results showed that the GEPT and the Cambridge English tests at the B1 level stimulated careful reading (Choice E in Item 1) the most frequently at the Whole Group level, and the operation 'understanding ideas which are not explicitly stated' (Choice D in Item 1) was performed the least frequently. Both tests required text-level comprehension (Choice C in Item 2) the most often (see shaded cells in Table 40).

Table 40 Wilcoxon signed-ranks test on the frequencies of pairs of cognitive skills used by different proficiency groups – the CEFR B1 level tests

Test	Item	Rank	High Group			Low Group			Whole Group		
			Choice	*z*	*p*	Choice	*z*	*p*	Choice	*z*	*p*
GEPT Intermediate	Cognitive processing	1	–	–	–	E. Careful reading	−2.20	.280	E. Careful reading	−4.48	.000*
		2	–	–	–	B. Search reading	−0.31	.755	C. From explicitly stated info	−0.76	.450
		3	–	–	–	D. From implicit info	−0.17	.858	B. Search reading	−0.27	.786
		4	–	–	–	C. From explicitly stated info	−0.15	.875	A. Scanning	−0.15	.878
		5	–	–	–	A. Scanning	–	–	D. From implicit info	–	–
	Level of comprehension	1	–	–	–	C. Text-level	−2.12	.030*	C. Text-level	−4.40	.000*
		2	–	–	–	B. Inter-sentential	−1.33	.182	B. Inter-sentential	−0.15	.878
		3	–	–	–	A. Intra-sentential	–	–	A. Intra-sentential	–	–
Cambridge English PET	Cognitive processing	1	C. From explicitly stated info	−1.43	.152	E. Careful reading	−2.20	.028*	E. Careful reading	−3.11	.002*
		2	A. Scanning	−1.20	.229	C. From explicitly stated info	−0.60	.550	B. Search reading	−2.82	.005*
		3	B. Search reading	−1.17	.240	D. From implicit info	−0.47	.638	A. Scanning	−1.07	.286
		4	E. Careful reading	−0.51	.609	B. Search reading	−0.01	.919	C. From explicitly stated info	−0.33	.745
		5	D. From implicit info	–	–	A. Scanning	–	–	D. From implicit info	–	–
	Level of comprehension	1	–	–	–	C. Text-level	−2.58	.009*	C. Text-level	−4.60	.000*
		2	–	–	–	B. Inter-sentential	−2.50	.012	B. Inter-sentential	−2.18	.029*
		3	–	–	–	A. Intra-sentential	–	–	A. Intra-sentential	–	–

**Within group difference between two consecutive choices reach significance at $p<.05$.*

The Mann-Whitney test was performed to investigate whether the High and the Low Groups processed the reading tasks differently. The results showed no significant between-group difference overall; except that when they were taking the GEPT Intermediate reading tests, the High Group applied careful reading (Choice E in Item 1, Table 17 in Chapter 3 and also Appendix 1) significantly less often than the Low Group (see the shaded cell in Table 41).

Table 41 Mann-Whitney test on the frequencies of the cognitive skills used by the High and Low Groups – the CEFR B1 level tests

	Item		**High-Low Group**	
			z	*p*
GEPT Intermediate	Cognitive processing	A. Scanning	−0.077	0.44
		B. Search reading	−1.42	0.16
		C. From explicitly stated info	−1.82	0.07
		D. From implicit info	−0.81	0.42
		E. Careful reading	−2.25	0.02*
	Level of comprehension	A. Intra-sentential	−0.89	0.37
		B. Inter-sentential	−0.43	0.66
		C. Text-level	−1.66	0.10
Cambridge English PET	Cognitive processing	A. Scanning	−1.15	0.25
		B. Search reading	−1.73	0.08
		C. From explicitly stated info	−0.37	0.71
		D. From implicit info	−0.81	0.42
		E. Careful reading	−1.52	0.13
	Level of comprehension	A. Intra-sentential	−1.48	0.14
		B. Inter-sentential	−0.68	0.50
		C. Text-level	−0.65	0.52

*$p<.05$

To examine whether the examinees used different cognitive skills when they took the GEPT Intermediate and Cambridge English PET reading tests, the Mann-Whitney test was performed. No significant within-group difference was observed, suggesting that the test takers used the same cognitive operations to answer the two tests at the CEFR B1 level, except that the Cambridge English PET required significantly more search reading (Choice B in Item 1, Table 17 in Chapter 3 and also Appendix 1) than the GEPT Intermediate. For details, see Table 42.

Overall, both the GEPT Intermediate and Cambridge English PET reading tests activated similar cognitive operations when the test takers were taking the tests, based on results of the Wilcoxon signed-ranks test (see Table 40): the examinees applied careful reading the most often, and a higher-order operation ‘understanding ideas which are not explicitly stated’ was

Table 42 The Mann-Whitney test on the frequencies of cognitive skills used by different proficiency groups when they took the GEPT and Cambridge English reading tests at the B1 level

Choice		GEPT-Cambridge					
		High Group		Low Group		Whole Group	
		z	*p*	*z*	*p*	*z*	*p*
Cognitive processing	A. Scanning	−1.32	0.19	−0.23	0.82	−0.69	0.49
	B. Search reading	−0.30	0.77	−0.32	0.75	−2.30	0.02*
	C. From explicitly stated info	−1.42	0.16	−0.69	0.49	−0.87	0.38
	D. From implicit info	−0.33	0.74	−0.35	0.73	−0.01	0.99
	E. Careful reading	−0.95	0.34	−0.40	0.69	−0.54	0.59
Level of comprehension	A. Intra-sentential	−0.59	0.55	−0.49	0.62	−0.32	0.75
	B. Inter-sentential	−0.149	0.62	−1.50	0.13	−1.77	0.08
	C. Text-level	−0.95	0.34	−0.92	0.35	−0.12	0.90

*p<.05

performed the least frequently at the whole group level; this corresponded to the results from expert judgement that over 85% of the questions in both tests required comprehension of textually explicit information; see Figure 37.

Both the GEPT and Cambridge English reading tests at the B1 level required text-level comprehension most frequently based on results of Wilcoxon signed-ranks test (see Table 40), and the difference between text-level comprehension and the other lower-level operations reached significance, suggesting that overall the GEPT Intermediate and Cambridge English PET reading tests stimulated the test takers to process the tasks globally, which required higher-order thinking. This finding was contradictory to the expert judgement (see Figure 39): the experts considered that both GEPT and Cambridge English tests at the B1 level contained a high proportion (around 40–50%) of questions that required within-sentence comprehension, and fewer than 20% of the items involved comprehension at the whole text level. It was speculated that the experts' level of language proficiency was clearly different from the test takers, so they read the texts in a different way from the test takers; the lower proficiency readers may have needed to process more of the text than the experts to arrive at an answer. Therefore, the experts failed to conceive how test takers processed the text.

GEPT and Cambridge English reading tests at B2 level

To investigate what cognitive processing skills test takers used when they were taking the GEPT High-Intermediate and Cambridge English FCE tests, 73 examinees (Groups 3 and 4 in Table 8 in Chapter 3) took both the

GEPT High-Intermediate and Cambridge English FCE reading tests and, in addition, were requested to fill out the Cognitive Processing Checklist (see Table 17 in Chapter 3 and Appendix 1) immediately after they answered each comprehension question. They were first rank-ordered based on their GEPT High-Intermediate and Cambridge English FCE reading test scores, and the highest and the lowest 27% were identified as the High Group (those with high English reading ability) and the Low Group (those with low English reading ability), respectively. The examinees classified to the same group, i.e. either High or Low Group, based on the scores from the two tests were selected for analysis.

A total of 14 out of 73 examinees were selected. The mean scores of the eight examinees in the High Group were 100.75 and 92.40 and those of the six examinees in the Low Group were 61.83 and 43.50 on the GEPT High-Intermediate and Cambridge English FCE reading test, respectively. The t-test result showed significant difference (*p*=.00) in reading performance of the High Group and the Low Group on both the GEPT High-Intermediate and the Cambridge English FCE reading tests. For details, see Table 43.

Table 43 T-test statistics on the High Group and the Low Group of the two CEFR B2 level tests

		High Group (N=8)	**Low Group (N=6)**	**t (*p*)**
GEPT High-Intermediate	Mean	100.75	61.83	9.20
	SD	9.69	3.97	(0.000**)
Cambridge English FCE	Mean	92.40	43.50	8.48
	SD	11.44	9.48	(0.000**)

SD=Standard deviation

***p<.01*

The mean frequencies, proportions, and standard deviations of each choice of the two items were computed for the High Group, the Low Group, and the Whole Group; see Table 44. Overall, examinees most frequently applied careful reading (Choice E in Item 1, Table 17 in Chapter 3 and also Appendix 1) when they took the GEPT High-Intermediate and Cambridge English FCE tests. Most test takers reported that they needed to understand the whole texts or information across sentences in order to find the answers to the Cambridge English FCE questions (Choices C and B, respectively, in Item 2). Compared to the cognitive operations performed during the FCE tests, the High Group reported that the GEPT High-Intermediate test involved more within-sentence comprehension (Choice A in Item 2). Moreover, it appeared they used search reading (Choice B in Item 1) more often when they took the Cambridge English FCE than when they took the GEPT High-Intermediate.

Table 44 Statistics on the mean frequencies, proportions, and standard deviations of each choice of the two items – the reading tests at the CEFR B2 level

			High Group		Low Group		Total	
			Mean (%)	SD	Mean (%)	SD	Mean (%)	SD
GEPT High-Intermediate	Cognitive processing	A. Scanning	11.50 (26%)	8.99	9.83 (22%)	9.06	10.15 (23%)	8.96
		B. Search reading	6.13 (14%)	4.12	16.33 (36%)	10.37	10.71 (24%)	8.08
		C. From explicitly stated info	11.63 (26%)	11.72	10.50 (23%)	6.80	10.74 (24%)	8.35
		D. From implicit info	6.25 (14%)	5.95	9.83 (22%)	6.82	10.15 (23%)	9.43
		E. Careful reading	24.00 (53%)	17.16	24.67 (44%)	12.85	20.25 (45%)	12.78
	Level of comprehension	A. Intra-sentential	20.63 (46%)	8.28	6.83 (15%)	7.94	16.52 (37%)	9.86
		B. Inter-sentential	10.38 (23%)	5.50	13.67 (30%)	11.93	11.37 (25%)	7.26
		C. Text-level	14.13 (31%)	10.33	24.67 (55%)	15.13	17.71 (39%)	11.47
Cambridge English FCE	Cognitive processing	A. Scanning	7.75 (26%)	6.63	9.83 (33%)	5.12	7.81 (26%)	7.47
		B. Search reading	9.63 (32%)	9.64	15.17 (51%)	4.22	12.10 (40%)	7.85
		C. From explicitly stated info	10.50 (35%)	10.85	13.67 (46%)	10.15	12.38 (41%)	8.13
		D. From implicit info	7.38 (25%)	6.19	13.17 (44%)	7.14	10.48 (35%)	6.95
		E. Careful reading	18.63 (62%)	12.07	17.00 (57%)	7.29	15.99 (53%)	8.92
	Level of comprehension	A. Intra-sentential	6.75 (23%)	4.13	3.33 (11%)	2.34	5.52 (18%)	4.81
		B. Inter-sentential	11.88 (40%)	8.04	13.33 (44%)	5.50	11.68 (39%)	6.55
		C. Text-level	11.50 (38%)	10.58	13.67 (46%)	7.20	13.10 (44%)	8.42

Note: the highest frequency choice within the group are marked in the shaded cells.
SD=Standard deviation

To examine whether differences in the frequencies among the choices of each item reach significance level within each group, the Friedman test was performed. Significant within-group difference (*p*<.05) was observed for cognitive processing (Item 1, Table 17 in Chapter 3 and Appendix 1) and level of comprehension (Item 2) when test takers were taking both the GEPT High-Intermediate and Cambridge English FCE at the Whole Group level. For details, see Table 45.

Table 45 Friedman test on the frequencies of the cognitive skills used by different proficiency groups – the CEFR B2 level tests

	Item	Group	df	N	X^2	*p*
GEPT High-Intermediate	Cognitive processing	Whole	4	73	26.54	.000*
		High	4	8	5.39	.249
		Low	4	6	8.65	.070
	Level of comprehension	Whole	2	73	7.27	.026*
		High	2	8	6.47	.039*
		Low	2	6	7.00	.030*
Cambridge English FCE	Cognitive processing	Whole	4	73	37.67	.000*
		High	4	8	2.29	.683
		Low	4	6	5.84	.211
	Level of comprehension	Whole	2	73	29.30	.000*
		High	2	8	2.64	.267
		Low	2	6	5.30	.070

**p<.05*

The Wilcoxon signed-ranks test was then performed to test within-group differences in frequencies that reached significance based on the Friedman test; see Table 46. The frequencies of choices for cognitive processing (Item 1, Table 17 in Chapter 3 and also Appendix 1) at the Whole Group level and of those for level of comprehension (Item 2) at the High, Low, and Whole Group levels were first rank-ordered, and then the difference in the use of two consecutive categories was tested. The alpha level was adjusted to 0.0125 for Item 1 and to 0.025 for Item 2, according to the Dunn-Bonferroni correction. Based on the Wilcoxon signed-ranks tests, the patterns of their use of cognitive skills were the same when they took the GEPT High-Intermediate and the Cambridge English FCE tests: the test takers reported applying careful reading most often, and the skill they used the least often was scanning. For level of comprehension (Item 2, Table 17 in Chapter 3 and also Appendix 1), test takers reported that they needed text-level comprehension to find the answers, but it appeared that the GEPT High-Intermediate tasks required test takers to process the texts less globally than the Cambridge English FCE tasks did.

The Mann-Whitney test was performed to investigate whether the High and the Low Groups processed the reading tasks differently; see Table 47.

Table 46 Wilcoxon signed-ranks test on the frequencies of pairs of cognitive skills used by different proficiency groups – the CEFR B2 level tests

Test	Item	Rank	High Group			Low Group			Whole Group		
			Choice	*z*	*p*	Choice	*z*	*p*	Choice	*z*	*p*
GEPT High-Intermediate	Cognitive processing	1	–	–	–	–	–	–	E. Careful reading	−4.27	0.000*
		2	–	–	–	–	–	–	D. From implicit info	−4.00	0.000*
		3	–	–	–	–	–	–	B. Search reading	−0.80	0.421
		4	–	–	–	–	–	–	C. From explicitly stated info	−0.16	0.876
		5	–	–	–	–	–	–	A. Scanning	–	–
	Level of comprehension	1	A. Intra-sentential	−1.26	0.207	B. Inter-sentential	−1.26	0.207	A. Intra-sentential	−3.60	0.000*
		2	C. Text-level	−0.70	0.484	C. Text-level	−0.94	0.345	C. Text-level	−0.19	0.849
		3	B. Inter-sentential	–	–	A. Intra-sentential	–	–	B. Inter-sentential	–	–
Cambridge English FCE	Cognitive processing	1	–	–	–	–	–	–	D. From implicit info	−5.33	0.000*
		2	–	–	–	–	–	–	E. Careful reading	−2.37	0.018*
		3	–	–	–	–	–	–	B. Search reading	−1.61	0.107
		4	–	–	–	–	–	–	C. From explicitly stated info	−1.01	0.311
		5	–	–	–	–	–	–	A. Scanning	–	–
	Level of comprehension	1	–	–	–	–	–	–	B. Inter-sentential	−5.62	0.000*
		2	–	–	–	–	–	–	C. Text-level	−0.49	0.623
		3	–	–	–	–	–	–	A. Intra-sentential	–	–

**Within group difference between two consecutive choices reach significance at $p<.05$.*

Table 47 The Mann-Whitney test on the frequencies of the cognitive skills used by the High and Low Groups – the CEFR B2 level tests

	Item		High-Low Group	
			z	*p*
GEPT High-Intermediate	Cognitive processing	A. Scanning	–0.45	0.65
		B. Search reading	−2.07	0.04*
		C. From explicitly stated info	0.00	1.00
		D. From implicit info	−1.17	0.24
		E. Careful reading	−0.19	0.85
	Level of comprehension	A. Intra-sentential	−2.45	0.01*
		B. Inter-sentential	−0.26	0.80
		C. Text-level	−1.62	0.11
Cambridge FCE	Cognitive processing	A. Scanning	−0.71	0.48
		B. Search reading	−0.91	0.36
		C. From explicitly stated info	−0.71	0.48
		D. From implicit info	−1.43	0.15
		E. Careful reading	−0.65	0.52
	Level of comprehension	A. Intra-sentential	−1.57	0.12
		B. Inter-sentential	−0.26	0.85
		C. Text-level	−1.10	0.27

**p<.05*

Between-group difference did not reach significance for Cambridge English FCE. However, when taking the GEPT High-Intermediate reading test, the High Group applied search reading significantly less frequently and within-sentence comprehension more often than the Low Group.

To examine whether there was significant difference in the frequencies of cognitive skills that the examinees used when they took the GEPT High-Intermediate and Cambridge English FCE reading tests, the Mann-Whitney test was again performed; see Table 48. As regards cognitive processing (Item 1, Table 17 in Chapter 3 and also Appendix 1), significant differences ($p<.05$) were observed when they took the two tests at the Whole Group level: they applied more careful reading, when they took the GEPT High-Intermediate than when they took the Cambridge English FCE, but they used more inferencing and search reading when they took the Cambridge English FCE than the GEPT High-Intermediate. As to levels of comprehension (Item 2, Table 17 in Chapter 3 and Appendix 1), the GEPT High-Intermediate tasks required more within-sentence understanding than the Cambridge English FCE tasks at the Whole Group and also at the High Group level, while the Cambridge English tasks required more across-sentence understanding than the GEPT tasks at the Whole Group level.

Overall, the GEPT and Cambridge English reading tests at the B2 level

Table 48 The Mann-Whitney test on the frequencies of cognitive skills used by different proficiency groups when they took the GEPT and Cambridge English reading tests at the B2 level

Choice		GEPT-Cambridge English					
		High Group		Low Group		Whole Group	
		z	*p*	*z*	*p*	*z*	*p*
Cognitive processing	A. Scanning	0.00	1.00	−1.44	0.15	−0.41	0.68
	B. Search reading	−0.95	0.34	−1.45	0.15	−4.27	0.00*
	C. Explicitly stated info	−0.37	0.71	−1.13	0.26	−4.26	0.00*
	D. From implicit info	−1.05	0.29	−1.60	0.11	−3.57	0.00*
	E. Careful reading	−0.58	0.56	−0.48	0.63	−2.26	0.02*
Level of comprehension	A. Intra-sentential	−2.21	0.03*	−0.24	0.81	−5.05	0.00*
	B. Inter-sentential	−1.26	0.21	−1.20	0.23	−4.54	0.00*
	C. Text-level	−0.63	0.53	−0.16	0.87	−1.44	0.15

**p<.05*

activated very different cognitive operations when the test takers were taking the tests. The examinees reported to apply careful reading the most often when they were taking the GEPT High-Intermediate reading test, while they needed to ‘understand ideas which are not explicitly stated’ most often to find the answers to the questions when they were taking the Cambridge English FCE test (see Table 46). Similarly, results from the Mann-Whitney test showed significant difference (*p*<.05) between frequencies of cognitive skills that the examinees used when they took the two tests at the Whole Group level. As regards cognitive processing, they applied significantly more careful reading when they took the GEPT High-Intermediate than when they took the Cambridge English FCE, but they used significantly more inferencing and search reading when they took the Cambridge English FCE than the GEPT High-Intermediate; see Table 48. Regarding levels of comprehension, the GEPT High-Intermediate required intra-sentential level comprehension the most frequently, while the FCE stimulated inter-sentential comprehension most often; see Table 46. The findings suggested that the Cambridge English FCE tasks activated higher-order cognitive operations, and GEPT High-Intermediate tasks stimulated the test takers to process the tasks more locally than the Cambridge English FCE, which might explain why the examinees scored higher on the GEPT High-Intermediate than on the Cambridge English FCE although textual features of the GEPT High-Intermediate reading tests were more complex than the FCE; see Table 35.

Discussion

The GEPT reading tests at CEFR B1 and B2 levels were compared with Cambridge English Reading tests at the B1 and B2 levels to assess whether the two CEFR-aligned reading tests at the same CEFR level were comparable in terms of test takers' performance, contextual features, and cognitive operations. In this study, core Cambridge English Reading papers at the B1 and B2 levels were selected as external criterion measures since they are among the few exams that have made claims about the relationship of their examinations to the levels of the CEFR, and other criterion measures used to validate CEFR-aligned relationships, such as DIALANG in Kecker and Eckes' (2010) study and exemplar tasks provided by Council of Europe (2005) in O'Sullivan's (2008) study, were not adequate for the intended purpose. The results showed that the GEPT and Cambridge English tests at the B1 level were comparable, while the GEPT and Cambridge English tests at the B2 level were not of the same difficulty.

As regards the GEPT and Cambridge English tests at the B1 level, no significant difference was observed between the test results (see Table 28), and they were comparable based on both contextual parameter analysis (see Table 34 and Figures 31 to 40) and cognitive processing analysis (see Table 42), except for a few features. For example, expert judgement on contextual features showed that the Cambridge English PET contained more varieties of genre than the GEPT Intermediate reading tests (see Figure 30), and automated textual analysis suggested that among those indices which were statistically different, the GEPT texts were closer to the norm in text length, mean number of words before main verb, and sentence syntax similarity, while Cambridge English texts were closer to Coh-Metrix Grade 12 norm texts in terms of lexical density, sentence length, and concreteness of the content words (see Table 36). Overall, the test takers processed the GEPT and Cambridge English tasks at the B1 level in a similar way (see Table 40), except that the Cambridge PET tasks required significantly more search reading that the GEPT Intermediate tasks (see Table 42).

On the other hand, the Cambridge English B2 level tests were significantly more difficult than the GEPT counterpart in terms of test takers' performance (see Table 31) and cognitive demands (see Table 48), but the Cambridge English texts were significantly less complex than the GEPT in terms of contextual features (see Table 35). Test takers scored significantly higher on the GEPT High-Intermediate reading test than on the FCE test (see Table 31), and both expert judgement and test takers' self-reports suggested that the FCE tasks were cognitively more challenging than the GEPT High-Intermediate tasks. Expert judges considered that the Cambridge English FCE tasks involved more higher-order cognitive processing, such as inferencing (see Figure 40), and contained more opinion questions (see Figure 38) and more questions that required understanding textually implicit information (see Figure 37) than the GEPT High-Intermediate tasks did. The test takers reported that they applied more inferencing and search reading when they took the Cambridge English FCE than the GEPT High-Intermediate (see Table 48). The Cambridge English texts at the B2 level were lexically and syntactically less complex than the GEPT counterpart, but they were conceptually less cohesive and contained significantly more pronouns than the GEPT texts at the B2 level (see Table 35).

Since the GEPT and Cambridge English B2 level texts were very different based on the automated textual analysis, the indices were compared with Coh-Metrix College level norm values (see Table 36). The results showed that almost all syntactic complexity, readability, and cohesion indices of the GEPT texts were closer to the norm values (see Table 36), except sentence syntax similarity and one of the text abstractness indices (i.e. Concreteness, Mean for content words). It is speculated that the inclusion of only narrative texts (see Figure 31) made textual features of the Cambridge English FCE texts clearly distinct from the College level norm texts and the GEPT B2 texts. The Cambridge English FCE tasks appeared to increase difficulty through imposing higher cognitive demands, e.g. including more inferencing and opinion questions, without taking features of the authentic texts into consideration.

One of the goals of the study is to identify criterial features that are useful to explicitly differentiate adjacent levels of reading proficiency. Therefore, textual features and cognitive operations of both the GEPT and Cambridge English tests were compared between the B1 and B2 levels to investigate whether different testing systems share the same rationale for test designs to differentiate between different levels of reading tests. The results of the textual analysis suggested that different testing systems had a different rationale for test design, although both testing systems tended to use shorter and more concrete texts in the B1 level reading papers than in the B2 level papers (see Table 32). Besides text length and text abstractness, other syntactical and lexical features of the GEPT and Cambridge English texts also showed significant difference ($p<.05$) between the B1 and B2 levels. Overall, the GEPT texts at the B2 level were lexically more complex and more difficult than those at the B1 level based on the various indices of lexical complexity, FRES and FK Grade Level indices; however, no significant difference was found between syntactic complexity and cohesion (see Table 25). On the other hand, the Cambridge English texts at the B1 and B2 levels were similar in terms of lexical and syntactical complexity, while, unexpectedly, the Cambridge English texts at the B2 level contained less variety of vocabulary (see STTR in Table 32), and were lexically less dense (see Lexical density indices) and more cohesive than those at the B1 level (see Referential cohesion indices), which suggested that the Cambridge English texts at the B2 level were easier in a number of respects than those at the B1 level (see Table 32).

Based on the results of the Coh-Metrix textual analysis (see Table 36), features of the GEPT reading texts at both the B1 and B2 levels were more similar to those of the real-life texts than their Cambridge English counterparts. Therefore, in terms of situational authenticity, the GEPT reading texts at both the B1 and B2 levels appeared to be appropriate for testing reading proficiency at the designated level, while the Cambridge English FCE texts might not be entirely representative in terms of textual features of the B2 level texts. That is to say, the GEPT tests at the B1 and B2 levels demonstrated higher context validity than the Cambridge English counterparts. On the other hand, the Cambridge English tests at the B1 and B2 levels demonstrated higher cognitive validity, since they stimulated test takers to use a wider variety of processing skills appropriate to levels than the GEPT tests did.

This study also demonstrates that expert judgement has its limitation, and there is a need to collect evidence of context and cognitive validity from various sources. Based on the results of textual analysis, it appeared that the experts tended to judge the tasks based on surface features of the texts intuitively. The automated tools could supplement expert judgement in providing a quantitative approach to comparing features of the texts used in the tests with those of authentic texts. More explicit information on characteristics of suitable texts could thus be provided to reflect test construction and to stabilise proficiency levels of a level-based test. As regards cognitive processing analysis, results from test takers' self-reports did not necessarily agree with those from expert judgement. It appeared that experts and test takers agreed on cognitive processing operations elicited during the tests (e.g. the Cambridge English tests at the B2 level required 'inferencing' more frequently than the GEPT counterpart, and the GEPT and Cambridge English tests at the B1 level required test takers to understand ideas which were not explicitly stated the least often), but the experts failed to predict the amount of text that test takers needed to process in order to find the answers to the questions. For example, the test takers reported that both the GEPT and Cambridge English tests at the B1 level required text-level comprehension most frequently (see Table 40), while experts considered that the tests takers only needed to process locally in order to find the answers most of the time (see Figure 39). Therefore, there is a need for collecting cognitive validity evidence from both expert judges' and test takers' perspectives during the test development stage.

The results of the studies presented in Chapters 4 and 5 are further discussed on the basis of the literature review in Chapter 6 to identify implications and directions for future research.

6 Conclusions and implications

The final chapter first answers the two research questions about the vertical differentiation of different levels within the GEPT, and the horizontal comparability of the GEPT and Cambridge English examinations targeting the same proficiency level. The validity evidence collected through a methodology which drew on Weir's (2005a) socio-cognitive validation framework in the design of its instruments is presented. Subsequently, the chapter will discuss the implications of the findings for test theory, test design, CEFR alignment procedures, and curriculum and course design procedures. The chapter will conclude with discussion of the limitations of the study and recommendations for future research.

Answers to the research questions about vertical differentiation of different levels within the GEPT and comparability of the GEPT and Cambridge English examinations targeting the same proficiency level

Answers to Research Question 1: Findings on relationships between the GEPT reading tests at CEFR B1 and B2 levels

The first research question asked: Is a GEPT reading test designed to measure at CEFR B2 level more difficult than a GEPT reading test designed to measure at CEFR B1 level in terms of test results, contextual parameters, and cognitive processing skills? Results from the vertical linking, and contextual parameter and cognitive processing analysis showed that a GEPT reading test targeting CEFR B2 level is more difficult than a GEPT reading test targeting CEFR B1 level.

Vertical linking scores from different levels of the GEPT on a common scale

Scores from different levels of a level-based exam cannot be compared straightforwardly since they are not based on the same score scale and thus do not share the same score unit. In this study, scores from different levels of the GEPT were linked and placed on a common score scale, using IRT Rasch model estimation. In this way, the difficulty of the GEPT levels could be compared based on the same score unit, and the extent of differentiation across the GEPT levels could then be established empirically.

The common-item non-equivalent groups design was adopted; i.e. a set of representative common test items was administered to different groups of examinees who had demonstrated systematically different levels of proficiency. GEPT Elementary, Intermediate, and High-Intermediate level reading tasks were selected to form shortened versions of the GEPT operational tests. The tasks were then grouped into two testlets: Testlet 1 was composed of Elementary and Intermediate reading test items, and Testlet 2, Intermediate and High-Intermediate test items. The same Intermediate test items were embedded in both testlets as an internal anchor, functioning as a basis for the linkage. Testlet 1 was administered to a group consisting of Elementary and Intermediate level target examinees, and Testlet 2 was administered to the other group consisting of Intermediate and High-Intermediate level test takers. The results showed item difficulty parameter (*b*) means increased with the GEPT levels. The means of difficulty parameter (*b*) estimates of the Elementary, Intermediate, and High-Intermediate reading test items were –1.57, –0.01, and 1.25, respectively; see Table 21 in Chapter 4. The difference between the Elementary and Intermediate levels was around 1.56, which was slightly larger than the difference between the Intermediate and High-Intermediate levels, around 1.26. The pattern of increases, in terms of item difficulty, was relatively regular across levels; see Figures 2 and 3 in Chapter 4. In addition, linear relationships were found between the scores obtained from the GEPT operational reading tests and the ability (θ) estimates, and the curves did not intersect at any points; see Figure 5 in Chapter 4. The findings demonstrated ascending difficulty across the GEPT levels in terms of observed test taker performance.

Comparisons of GEPT test constructs between the B1 and B2 levels

Following the vertical linking to examine the relationships of the scores from different levels of the GEPT, the contextual features and cognitive operations activated during the tests were compared to look into the test constructs of the GEPT at the B1 and B2 levels and to identify criterial features that distinguished GEPT levels. Textual features that affected the comprehensibility and difficulty of the reading tasks were analysed using both traditional expert judgment and the automated tools; the automated textual analysis built on the procedures introduced by Green et al (2010). The results of the automated textual analysis showed significant difference in text length, text concreteness, lexical complexity, and readability indices between the two levels of the GEPT, while no significant difference in syntactic complexity or text cohesion was observed; see Table 25 in Chapter 4. In terms of text type, the GEPT Intermediate texts were mostly expository and the GEPT

High-Intermediate texts contained a wider variety of rhetorical organisations, including a rhetorically more demanding text type – argumentative texts; see Figure 8 in Chapter 4. Other features of the GEPT reading texts at the B1 and B2 levels were similar: most texts were in the social domain (see Figure 6), belonged to the magazine and newspaper article report genre (see Figure 7), and were explicitly organised (see Figure 9), semantically concrete (see Figure 10), and subject (see Figure 11) and cultural neutral (see Figure 12). Most of the test questions from GEPT reading papers at the B1 and B2 levels were specific detail and factual questions (see Figure 13), for which the test takers could find the answers based on explicitly stated information in the texts (see Figure 14). Test takers' cognitive operations activated during the tests were analysed using expert judgment as advocated by Khalifa and Weir (2009). The GEPT tests at the B1 level contained a higher proportion of questions that required within-sentence comprehension than those at the B2 level, while fewer than 20% of the items of both the GEPT at the B1 and B2 levels involved comprehension at the whole text level; see Figure 15 in Chapter 4.

The findings showed that a GEPT reading test designed to measure at CEFR B2 level was more difficult than a GEPT reading test designed to measure at CEFR B1 level in terms of both contextual features and cognitive operations.

Comparability of the GEPT and Cambridge English exams targeting the same proficiency level

The second research question asked: Are GEPT reading tests at CEFR B1 and B2 levels comparable to alternative CEFR-linked measures in terms of test results, contextual parameters, and cognitive processing skills? Core Cambridge English Reading papers at the B1 and B2 levels were selected as the external criterion measures in this study. The results showed that the GEPT and Cambridge English tests at the B1 level were comparable, while the GEPT and Cambridge English tests at the B2 level were not.

Comparisons of the GEPT and Cambridge English reading tests at the B1 level

The GEPT and Cambridge English tests at the B1 level were comparable overall. Regarding test takers' performance, no significant difference was observed between the test results of the GEPT and Cambridge English reading tests at the B1 level; see Table 28 in Chapter 5. The distributions of test takers' scores from the two tests were both symmetric; see Figure 25 in Chapter 5. In terms of contextual parameters, the GEPT and Cambridge English tests at the B1 level were comparable as well (see Table 34 and Figures 31 to 35), except for a few features. For example, the Cambridge English B1 texts contained more varieties of genre than the GEPT B1 reading texts; see Figure 30. The automated textual analysis suggested that among those indices which were statistically different, the GEPT texts were closer to Coh-Metrix Grade 12 norm texts in text length, mean number of words before main verb and sentence syntax similarity, while Cambridge English texts were closer to the norm in terms of lexical density, sentence length, and concreteness of the content words; see Table 36 in Chapter 5. It appeared that the GEPT reading texts at the B1 levels were more similar to those of the real-life texts than their Cambridge English counterparts. Therefore, in terms of situational authenticity, the GEPT appropriate reading texts were for testing reading at the B1 level.

As to cognitive operations, the test takers processed the GEPT and Cambridge English tasks in a similar way based on results from expert judgment (see Figures 36 to 40 in Chapter 5) and test takers' self-reports (see Table 42 in Chapter 5), except that the test takers reported that the Cambridge English B1 tasks required search reading significantly more often than the GEPT B1 tasks did (see the shaded area in Table 42). Nevertheless, results about scope of text processing from test takers' self-reports differed from those from expert judgment; while test takers reported that both the GEPT and the Cambridge English reading tests at the B1 level stimulated them to process the tasks globally, which required higher-order thinking (see Table 40), experts considered both the B1 level tasks of the two examinations stimulated test takers to process at intra-sentential level (see Figure 39), which required lower order thinking, the most often. It was speculated that the experts' level of language proficiency was clearly different from the test takers, so they read the texts in a different way from the test takers and, therefore, were unable to conceive how test takers processed the text.

Comparisons of the GEPT and Cambridge English reading tests at the B2 level

In terms of test takers' performance, the Cambridge English B2 level tests were significantly more difficult than the GEPT counterpart; see Table 31 in Chapter 5. Test takers scored significantly lower on the Cambridge English B2 Reading test than on the GEPT B2 test. The distribution of test takers' scores on the Cambridge English B2 Reading test was symmetric, while that of the GEPT B2 test was negatively skewed; see Figure 27. Moreover, the Cambridge English B2 Reading tasks were cognitively more challenging than the GEPT B2 tasks. Expert judges considered that the Cambridge English tasks contained more questions that required understanding textually implicit information (see Figure 37) and fewer factual and more opinion questions (see Figure 38), and involved more higher-order cognitive processing, such as inferencing (see Figure 40), than the GEPT B2 tasks did. Similarly, test takers' self-reports suggested that they applied inferencing and search reading more often when they took the Cambridge English test than the GEPT test; see Table 48.

Nevertheless, the GEPT texts were significantly more complex than the Cambridge English texts in terms of contextual features; see Table 35 in Chapter 5. The GEPT texts at the B2 level were lexically and syntactically more complex than the Cambridge English counterpart, but they were conceptually more cohesive and contained significantly fewer pronouns than the Cambridge English texts at the B2 level.

Based on the results of Coh-Metrix textual analysis (see Table 36 in Chapter 5), almost all syntactic complexity, readability, and cohesion indices showed that the features of the GEPT texts at the B2 levels were closer to those of the texts in Coh-Metrix College level norms than those of the Cambridge English counterpart. In terms of situational authenticity, the GEPT reading texts were appropriate for testing reading proficiency at the B2 level.

Implications for test theory

This study adopted the procedures proposed in the Manual (Council of Europe 2003, 2009) and Weir's (2005a) socio-cognitive validation framework to gather validity evidence of GEPT level differentiation and to make comparisons between the GEPT and Cambridge English tests at the B1 and B2 levels. To answer Research Question 1 (Is a GEPT reading test designed to measure at CEFR B2 level more difficult than a GEPT reading test designed to measure at CEFR B1 level in terms of test results, contextual parameters, and cognitive processing skills?) criterion-related validity evidence was generated through vertically linking scores from different levels of the GEPT, and contextual features and cognitive operations were compared between the GEPT at the B1 and B2 levels to provide context validity evidence and cognitive validity evidence, respectively. To answer Research Question 2 (Are GEPT reading tests at CEFR B1 and B2 levels comparable to alternative CEFR-linked measures in terms of test results, contextual parameters, and cognitive processing skills?) horizontal comparisons between the contextual features, cognitive operations and test results of the GEPT and Cambridge English reading tests at the same CEFR level were made to provide context, cognitive, and criterion-related validity evidence. The validation procedures for level differentiation proposed in this study are presented in Figure 41.

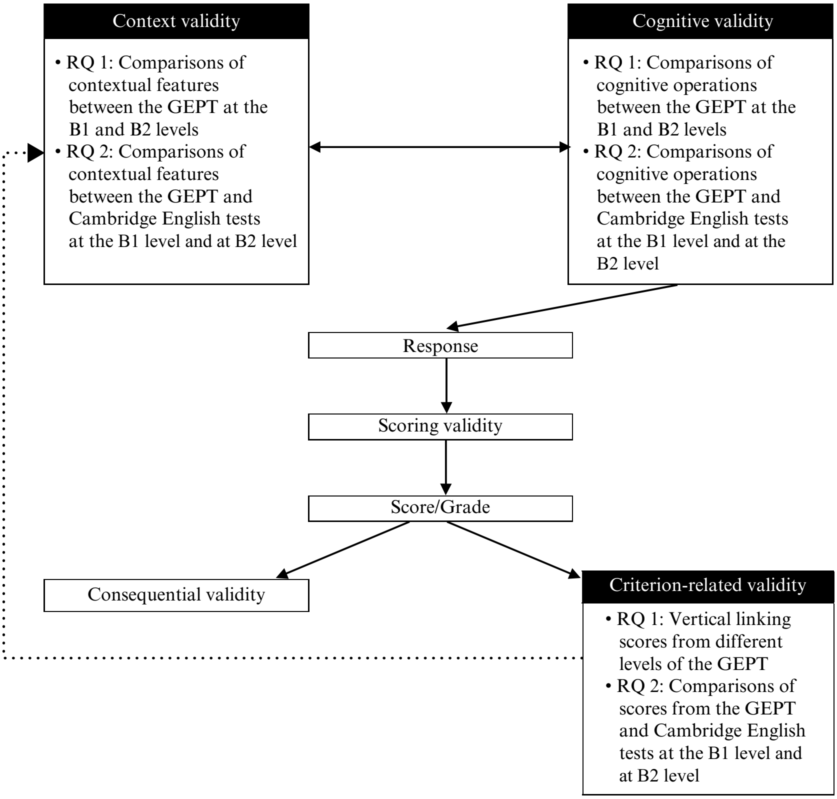
Various case studies (e.g. Kecker and Eckes 2010, O'Sullivan 2008, Wu and Wu 2010) reported that they found the Manual (2003) useful to relate the exams to the CEFR levels; nevertheless, O'Sullivan (2008:85) suggested that 'limiting the validation evidence to estimates of internal and external validity is far too simplistic a view of validation' and advocated collecting the validity evidence based on an explicit model of validation, such as Weir's (2005a) socio-cognitive validation framework, to provide theoretical justification behind the linking relationship (O'Sullivan 2008:51). Recent research (e.g. Shaw and Weir 2007, Khalifa and Weir 2009) has proved useful in amplifying Weir's (2005a) framework, thereby enabling researchers to generate comprehensive validity evidence on cognitive and contextual distinctions across different levels of proficiency.

This study also went beyond the scope of the earlier studies by providing transparent procedures to link scores from different test levels in order to statistically demonstrate the extent of overall differentiation across test levels, and suggested a comprehensive methodology for comparison of exams which are developed by different exam boards targeting the same proficiency levels.

This study added support to Weir and Taylor's (2011:234) contention that Weir's (2005a) socio-cognitive framework:

> . . . has direct relevance and value to an operational language testing/assessment context . . . [O]ther frameworks (e.g. Bachman 1990) were helpful in provoking us to think about key issues from a theoretical perspective but they generally proved very difficult for practitioners to operationalise in a manageable and meaningful way.

Figure 41 Validation procedures for GEPT level differentiation based on Weir's (2005b) socio-cognitive framework


Context validity
• RQ 1: Comparisons of contextual features between the GEPT at the B1 and B2 levels
• RQ 2: Comparisons of contextual features between the GEPT and Cambridge English tests at the B1 level and at B2 level
Cognitive validity
• RQ 1: Comparisons of cognitive operations between the GEPT at the B1 and B2 levels
• RQ 2: Comparisons of cognitive operations between the GEPT and Cambridge English tests at the B1 level and at the B2 level
Response
Scoring validity
Score/Grade
Consequential validity
Criterion-related validity
• RQ 1: Vertical linking scores from different levels of the GEPT
• RQ 2: Comparisons of scores from the GEPT and Cambridge English tests at the B1 level and at B2 level

The framework recognises the significance of collecting various aspects of validity at different phases of the testing cycle in a temporal sequence (O'Sullivan and Weir 2011). It would, however, be useful to establish a feedback loop (see the dotted line with arrow ending in Figure 41) from criterion-related validity to context validity so that validation results can contribute to ongoing improvement of test quality.

Implications for test design

Bachman and Palmer (1996) argued that situational and interactional authenticities were essential features of useful test tasks. Although full 'situational authenticity' was generally not achievable within the constraints of testing conditions, Khalifa and Weir (2009:81) argued that 'the contextual parameters operationalised in a test should mirror the criterial features of the target situation activity as far as is possible.' Test developers traditionally rely on expert judgement to derive holistic interpretations of test specifications which outline the rationale of test design and are intended to enable test content and difficulty to remain consistent across different test forms. Recent advances in computational linguistics and the development of corpora provide test developers with a workable and efficient approach to automating analysis of a range of individual textual characteristics (Green et al 2010). Automated textual analysis can supplement expert judgement to define contextual features of the tests more explicitly than the traditional approach, and also provide a quantitative method for comparing features of the texts used in the tests with those of authentic texts, and thus allow test developers to determine whether the tasks in the tests reflect real-world reading activities.

This study demonstrated the feasibility of collecting context validity evidence using automated tools. It offered exam boards a manageable approach for establishing the extent to which reading reflects accepted test constructs and for stablising proficiency levels of their level-based tests. In this study, indices obtained from the automated textual analysis of the test tasks were compared with Coh-Metrix norm values which were computed based on sample texts from TASA corpus (Landauer et al 1998). Results from the automated textual analysis showed that textual features of the GEPT reading tasks at both the B1 and B2 levels were closer to the real-life sample texts from the TASA corpus than their Cambridge English counterparts. Overall, the GEPT reading texts at both the B1 and B2 levels were appropriate for testing reading proficiency at the designated level, in terms of situational authenticity.

The findings of this study suggested several modifications to further enhance the features of the GEPT texts to more closely represent real-life reading tasks (see Table 36 in Chapter 5), including:

1. Extending the length of sentences in both the B1 and B2 texts: it appeared that average lengths of sentences at both the B1 and B2 levels (15.66 and 19.23 words, respectively) were shorter than those of the norms (19.63 and 22.03 words, respectively).
2. Increasing degree of text abstractness at both the B1 and B2 levels: Coh-Metrix indices of concreteness for content words suggested that the GEPT at both the B1 and B2 levels (413.33 and 394.07, respectively) were more concrete than the norms (377.52 and 372.54, respectively).
3. Using a wider variety of sentence structures in the B2 texts: based on sentence syntax similarity index, difference between the GEPT B1 and B2 texts (0.110 and 0.112, respectively) was not significant at the 0.05 level. The range of sentence structures used in the B1 texts was very close to Grade 12 Norm texts (0.10), while the B2 texts used a narrower range of sentence structures than the College level norm texts (0.09), suggesting that adjusting the variety of sentence structures would make the GEPT B2 texts closer to the real-world reading texts.

As for Cambridge English texts, textual features of the Cambridge English B1 texts (i.e. PET) were comparable to those of the Grade 12 norm texts, except that mean number of words before main verb of main clauses in the Cambridge English B1 texts was 2.56 words, considerably fewer than the norm texts (i.e. 5.20 words). However, the Cambridge English B2 (i.e. FCE) texts appeared to be very different from the College level norm texts, according to almost all indices from the automated textual analysis. The inclusion of only narrative texts might have contributed to making textual features of the Cambridge English B2 Reading texts clearly distinct from the College level norm texts: the Cambridge English B2 texts were lexically and syntactically less complex, but conceptually less cohesive, and they contained significantly more pronouns. Therefore, it is speculated that the Cambridge English FCE might not be wholly representative of CEFR B2 level tests in terms of textual features.

When it comes to task authenticity, in addition to contextual parameters, it was also important to approximate to the cognitive operations elicited in the real world when candidates read test tasks (Alderson 2000:53, Green et al 2010:2). To investigate the cognitive operations required for test takers to process the test tasks, this study collected evidence from both expert judges' and examinees' perspectives, and the results from both sources were triangulated. Results from test takers' self-reports did not agree with those from expert judgment; this corresponded to the results from previous studies (e.g. Weir, Hawkey, Green, Devi and Ünaldi 2009). In this study, experts considered that both the GEPT and Cambridge English tests at the B1 level required sentence-level comprehension the most frequently, while test takers reported that to find the answers they needed text-level comprehension the most frequently, and the difference between text-level and the other lower-level operations reached significance level ($p<.05$). This study demonstrated the need for collecting *a priori* cognitive validity evidence from the test takers' perspective. The discrepancies between results from the expert judgment and test takers' self-reports suggested that expert judgment alone was not sufficient, and piloting with test takers during the test development stage to establish how the examinees actually process test tasks under test conditions was needed.

The Cambridge English tests at the B1 and B2 levels incorporate expeditious reading, i.e. Parts 2 and 3 in the B1 tests, and Part 3 in the B2 tests, while the GEPT tests at both levels assess careful reading only. The results showed that the B1 level test takers applied search reading significantly ($p<.05$) more often when they took the Cambridge English test than they did during the GEPT test; see Table 42. Similarly, the B2 level test takers applied both search reading and inferencing significantly ($p<.05$) more often when they took the Cambridge English test than they did during the GEPT test; see Table 48. These suggested although the Cambridge English B2 texts might not be wholly representative in terms of textual features of B2 levels texts, they stimulated test takers to use a wider variety of processing skills than the GEPT B2 tasks did, and thus they demonstrated higher cognitive validity. Further investigation into possible revision of the GEPT at the B1 and B2 level is necessary to reflect more closely the real-life reading processes.

In addition, results of the cognitive processing and contextual parameter analysis in this study suggested cognitive operations elicited when test takers were taking the tests appeared to exert a marked influence on reading comprehension difficulty, while surface textual features, such as word frequency, sentence length, and syllables per words, which are traditionally considered to have significant impact on reading comprehension, did not necessarily play a decisive role in task difficulty. This finding corresponded to Alderson's (2000: 70) views that 'at some level the syntax and lexis of texts will contribute to text and thus test difficulty, but the interaction among syntactic, lexical, discourse and topic variables is such that no one variable can be shown to be paramount.'

Automated textual analysis is of value not only in construct validation but also in item writer training as it provides more explicit information on characteristics of suitable texts to help item writers develop texts which resemble authentic real-life reading tasks. Therefore, it is suggested the methodology in this study be incorporated into their regular item development practice. Nevertheless, caution should be exercised in using the indices produced by Coh-Metrix, as the limited sample size of some of the corpora and also the sources of the texts in the corpora, e.g. the MRC Psycholinguistics Database (Coltheart 1981) and TASA corpus (Landauer et al 1998), which Coh-Metrix use to compute the indices, may not be fully representative.

Implications for CEFR alignment procedures

To support test providers in relating their examinations to the CEFR levels and validating the linking relationship, the Council of Europe published *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*: *A Manual* (Council of Europe 2003, 2009). The Manual (Council of Europe 2009) proposed five inter-related sets of procedures, i.e. Familiarisation, Specification, Training/Standardisation, Standard Setting, and Validation, for institutions that wish 'to make claims about the relationship of their examinations to the levels of the CEFR' to design a linking scheme and 'to demonstrate the validity of those claims' (2009:2). Kecker and Eckes (2010) and O'Sullivan (2008) considered the methodology's provision of a pragmatic approach for the alignment purpose. At the same time, both studies reported problems with descriptors in the CEFR scales; e.g. some parts of descriptors were not applicable to or observed in the exams, while some examination content and the notion of task fulfilment were not covered in the CEFR scales (Kecker and Eckes 2010). Other researchers (Alderson 2007, Figueras and Noijons 2009, Martyniuk and Noijons 2007, Morrow 2004, Weir 2005a) also called for parameters such as structures, lexis or other linguistic features, in the CEFR Can Do scales to be more explicitly defined and to provide clearer guidance on relating tasks to a specific level in terms of these (e.g. Alderson et al 2006:13, Weir 2005b:292). In addition to vague level descriptors in relation to contextual features, Alderson (2007:661) noted that 'there was no theory of comprehension that could be used to identify the mental operations that a reader or listener has to engage in at the different levels of the CEFR.' This study provided empirical procedures to collect evidence on cognitive validity using both expert judgement and test takers' self-reports in a manageable way.

Implication for curriculum and course design procedures

The CEFR is designed for language learning, teaching, and assessment. Nevertheless, the CEFR has considerably less significance for and impact on teaching and learning than on testing. As Little (2007:648) noted '[t]o date, its impact on language testing far outweighs its impact on curriculum design and pedagogy.' A few examples of the CEFR being used in language teaching and learning he presented included:

1. Relating learning outcomes to the CEFR levels in France: secondary students are expected to achieve the B1 level in their first and the A2 level in their second foreign language by the end of compulsory education.
2. Developing curricula, e.g. for adult language learners in Catalonia (Figueras and Melcion 2002), and learning supports, e.g. for ESL learners

in Irish primary schools (Little 2005, Little and Lazenby Simpson 2004), based on the CEFR's descriptive scheme (Little 2007: 649).

Westhoff (2007:676) argued that 'although the CEFR descriptors tell us a lot about what learners at a certain level can do, very little is stated about what they should know in order to carry out these language tasks', which is essential to set curricular objectives.

Urquhart and Weir (1998:172) recognised similarities between testing and teaching in determining reading activities, in which appropriate texts need to be selected for readers to perform activities developed for them under appropriate performance conditions. Grabe (2009) suggested cognitive operations elicited were affected more by the difficulty of the text than proficiency levels of the readers. Grabe (2009:228) explained that:

> A normally good reader who is reading a frustration-level text (usually when readers know fewer than 90 percent of the words in the text) will engage in more 'local' strategies and behave like a poor reader with respect to strategy use.

It is therefore equally important for teaching institutions and course designers to identify salient features of reading texts which are appropriate to learners' proficiency levels. Procedures for analysing textual features and identifying cognitive processes in this study provided a workable approach for them to operationalise reading constructs that better reflect real-life reading activities when they prepare reading tasks for language learning.

Research limitations and suggestions for future research

The generalisability of the results obtained in this study on the validation of different levels of the GEPT reading tests in terms of test constructs, i.e. contextual features and cognitive operations stimulated during the tests, and test results is limited since this study is of restricted scope. There are several limitations to the inferences that should be addressed in future research.

The vertical linking study on differentiation of the GEPT levels in terms of difficultly was based on a small sample and the use of Rasch model calibrations only. Skaggs and Lissitz (1985) suggested that Rasch model estimation for vertical scaling might produce inconsistent results. If more data were collected, more refined statistical methods (e.g. two- or three-parameter IRT model) could be applied to produce parameter estimates that were more stable. Furthermore, this study involved only the tests that are dichotomously scored since BILOG-MG, the software used to scale parameter estimates in this study, can only apply to models for dichotomous items. It will be useful in future studies to apply more sophisticated software, e.g.

MULTILOG, to link test items that are not dichotomously scored onto a common score scale.

The studies on contextual features and cognitive operations which distinguished between adjacent levels of reading proficiency used only small samples, and the test takers' background was homogeneous. Limited test papers were available for analysis; for contextual features, only three papers from each exam were analysed. In addition, the effects of familiarity of the test formats on test takers' performance cannot be overlooked. In this study, all test takers were familiar with the GEPT test formats, but they had had little exposure to the Cambridge English exams prior to this study and were briefed on the Cambridge English tests only before taking the tests. Therefore, the interactions between the test tasks and the test takers need to be further investigated.

Moreover, the impact of test characteristics, such as time constraints, text length, and item types, has not yet been established. Firstly, comprehension should be assessed with reference to response time. Alderson (2000:30) suggested that 'speed should not be measured without reference to comprehension'. Khalifa and Weir (2009:100) argued 'when candidates are given a short text, they may tend to employ a careful bottom up approach to reading rather than expeditious if the time has not been constrained to prevent this'. In addition, the tests involved in this study consisted of MCQs only. Embretson and Wetzel's (1987) model for responding to MC reading questions suggested that MC questions create very different comprehension and response processes. Rupp et al (2006:441) argued that 'learners view responding to multiple-choice questions as a problem-solving task rather than a comprehension task.' Therefore, it is suspected that assessing reading comprehension with MC questions might change reading processes and activate particular response processes as a result. Future studies are advised to include test takers who have a greater variety of background to investigate differences between cognitive operations elicited when they are responding to MC items and other item types, such as short answer questions, under different time constraints.

In this study, test takers' cognitive processes are identified through expert judgment using the Cognitive Processing Proforma (see Appendix 3) and test takers' self-report using the Cognitive Processing Checklist (see Table 17 in Chapter 3 and Appendix 1). Very limited informants, only 12 expert judges and 144 examinees, completed questionnaires on cognitive processes; more data should be collected to confirm whether the results observed in this study can generalise to a larger population. Future studies are also advised to adopt a variety of approaches for collecting introspective data, such as planned interviews, prompted retrospections or eye tracking, to triangulate the results, and also to investigate the relative value of expert judgment and the procedures mentioned above.

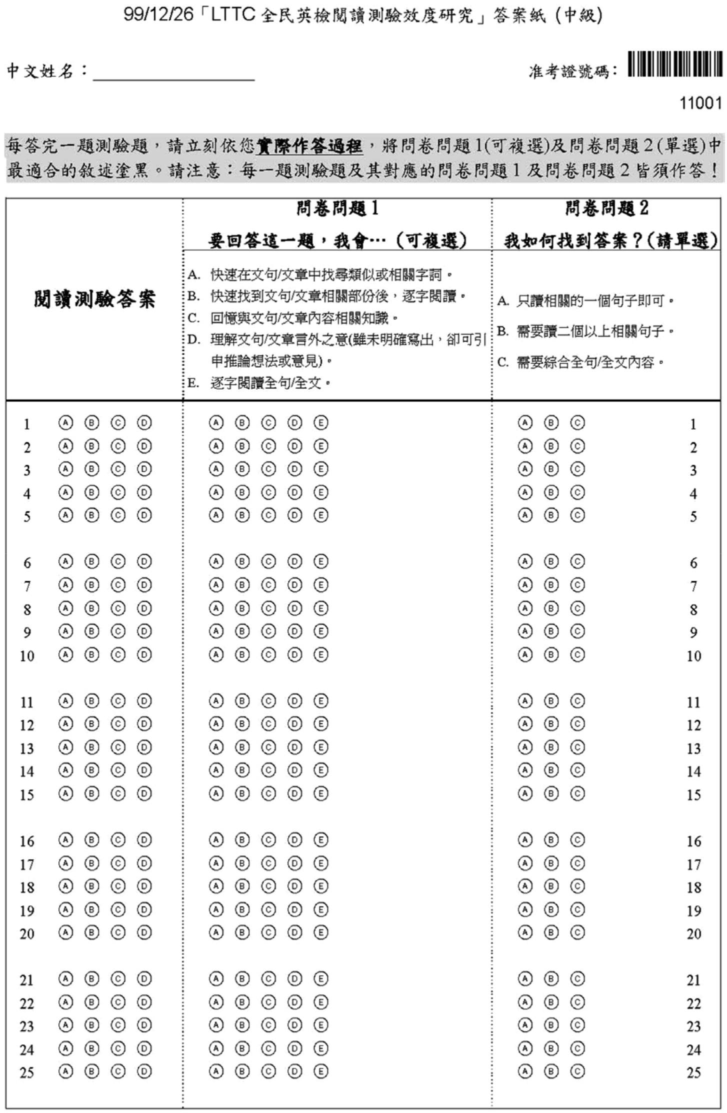
Conclusions

This study utilised Weir's (2005a) socio-cognitive validation framework to develop an innovative set of methodological procedures for examining various aspects of the validity of different levels of the GEPT in terms of test results, contextual parameters, and cognitive processing skills. The CEFR and two levels of a CEFR-aligned multilevel test battery, PET and FCE developed by Cambridge English Language Assessment, served as external referents for a review of the similarities and differences between the GEPT reading tests targeting CEFR B1 and B2 levels. To establish 'situational and interactional authenticities' (Bachman and Palmer 1996), this study not only applied automated tools and expert judgment to examine 'the degree of correspondence of the characteristics of a given language test task to the features of a TLU task' (1996:23), but also successfully realised with clear results what O'Sullivan (2006:183) advocates, namely 'an *a posteriori* empirical exploration of test performance' to gather evidence of interactional authenticity. The findings supported the construct validity of the GEPT in general, but showed that its cognitive validity needs to be enhanced by incorporating tasks that test expeditious reading operations and inferencing at the B2 level.

Finally, the study sends a timely warning to those who simplistically believe that using the procedures that the Manual (Council of Europe 2003, 2009) recommends to link an examination to a CEFR level can demonstrate an equivalence with other examinations that have been located at this particular CEFR level. In this study, by employing a far more sophisticated set of procedures for comparison, we have demonstrated that my B1 might be the same as your B1 in many respects, but my B2 does not seem to be your B2.

Appendix 1 Cognitive Processing Checklist

An English language version of this checklist can be found in Table 17, Chapter 3.



99/12/26「LTTC 全民英檢閱讀測驗效度研究」答案紙 (中級)

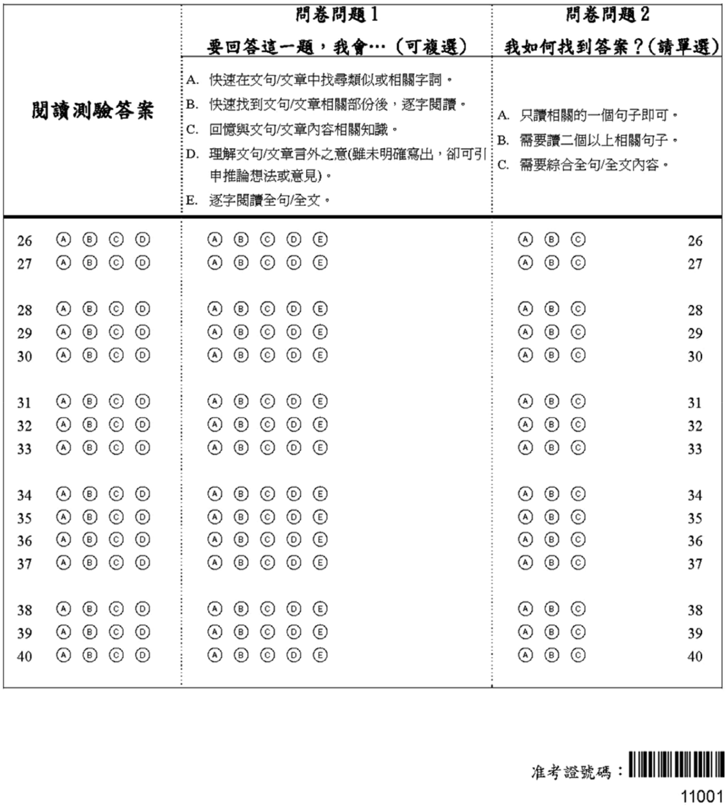
中文姓名：________________

准考證號碼：

11001

每答完一題測驗題，請立刻依您**實際作答過程**，將問卷問題1(可複選)及問卷問題2(單選)中最適合的敘述塗黑。請注意：每一題測驗題及其對應的問卷問題1及問卷問題2皆須作答！

	閱讀測驗答案	問卷問題1 要回答這一題，我會…（可複選）	問卷問題2 我如何找到答案？(請單選)	
		A. 快速在文句/文章中找尋類似或相關字詞。 B. 快速找到文句/文章相關部份後，逐字閱讀。 C. 回憶與文句/文章內容相關知識。 D. 理解文句/文章言外之意(雖未明確寫出，卻可引申推論想法或意見)。 E. 逐字閱讀全句/全文。	A. 只讀相關的一個句子即可。 B. 需要讀二個以上相關句子。 C. 需要綜合全句/全文內容。	
1	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	1
2	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	2
3	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	3
4	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	4
5	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	5
6	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	6
7	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	7
8	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	8
9	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	9
10	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	10
11	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	11
12	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	12
13	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	13
14	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	14
15	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	15
16	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	16
17	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	17
18	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	18
19	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	19
20	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	20
21	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	21
22	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	22
23	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	23
24	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	24
25	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	25

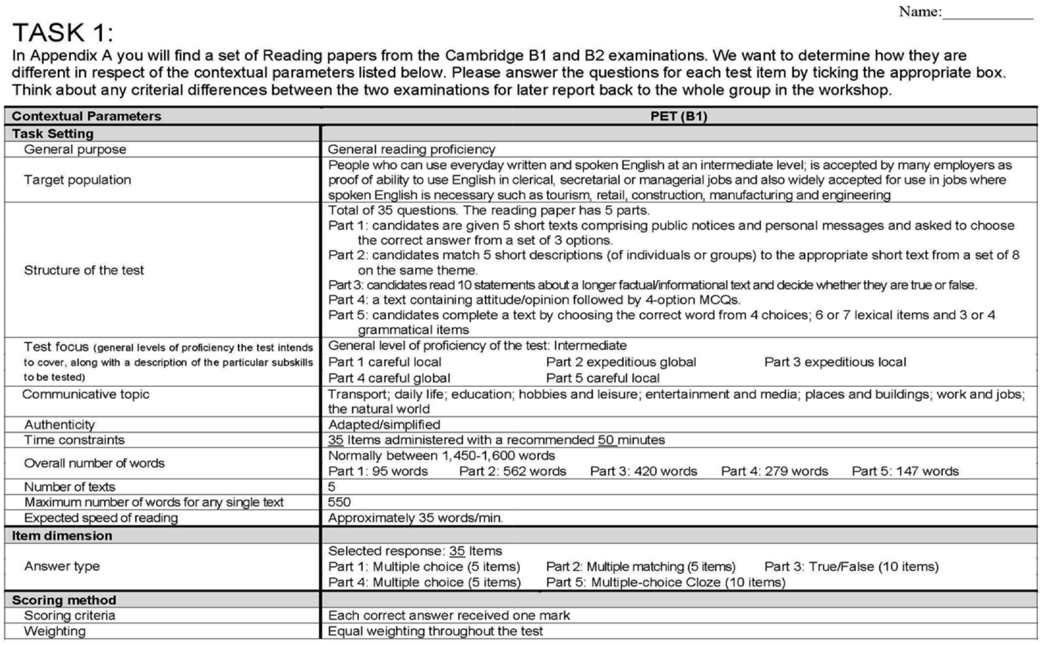


	閱讀測驗答案	問卷問題1 要回答這一題，我會…（可複選）	問卷問題2 我如何找到答案？(請單選)	
		A. 快速在文句/文章中找尋類似或相關字詞。 B. 快速找到文句/文章相關部份後，逐字閱讀。 C. 回憶與文句/文章內容相關知識。 D. 理解文句/文章言外之意(雖未明確寫出，卻可引申推論想法或意見)。 E. 逐字閱讀全句/全文。	A. 只讀相關的一個句子即可。 B. 需要讀二個以上相關句子。 C. 需要綜合全句/全文內容。	
26	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	26
27	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	27
28	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	28
29	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	29
30	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	30
31	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	31
32	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	32
33	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	33
34	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	34
35	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	35
36	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	36
37	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	37
38	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	38
39	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	39
40	Ⓐ Ⓑ Ⓒ Ⓓ	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ	Ⓐ Ⓑ Ⓒ	40

准考證號碼：
11001

Appendix 2 Contextual Parameter Proforma

The PET (B1) Contextual Parameter Proforma is provided below as an illustration. A very similar form was used for FCE (B2).



Name:___________

TASK 1:

In Appendix A you will find a set of Reading papers from the Cambridge B1 and B2 examinations. We want to determine how they are different in respect of the contextual parameters listed below. Please answer the questions for each test item by ticking the appropriate box. Think about any criterial differences between the two examinations for later report back to the whole group in the workshop.

Contextual Parameters	PET (B1)
Task Setting	
General purpose	General reading proficiency
Target population	People who can use everyday written and spoken English at an intermediate level; is accepted by many employers as proof of ability to use English in clerical, secretarial or managerial jobs and also widely accepted for use in jobs where spoken English is necessary such as tourism, retail, construction, manufacturing and engineering
Structure of the test	Total of 35 questions. The reading paper has 5 parts. Part 1: candidates are given 5 short texts comprising public notices and personal messages and asked to choose the correct answer from a set of 3 options. Part 2: candidates match 5 short descriptions (of individuals or groups) to the appropriate short text from a set of 8 on the same theme. Part 3: candidates read 10 statements about a longer factual/informational text and decide whether they are true or false. Part 4: a text containing attitude/opinion followed by 4-option MCQs. Part 5: candidates complete a text by choosing the correct word from 4 choices; 6 or 7 lexical items and 3 or 4 grammatical items
Test focus (general levels of proficiency the test intends to cover, along with a description of the particular subskills to be tested)	General level of proficiency of the test: Intermediate Part 1 careful local Part 2 expeditious global Part 3 expeditious local Part 4 careful global Part 5 careful local
Communicative topic	Transport; daily life; education; hobbies and leisure; entertainment and media; places and buildings; work and jobs; the natural world
Authenticity	Adapted/simplified
Time constraints	35 Items administered with a recommended 50 minutes
Overall number of words	Normally between 1,450-1,600 words Part 1: 95 words Part 2: 562 words Part 3: 420 words Part 4: 279 words Part 5: 147 words
Number of texts	5
Maximum number of words for any single text	550
Expected speed of reading	Approximately 35 words/min.
Item dimension	
Answer type	Selected response: 35 Items Part 1: Multiple choice (5 items) Part 2: Multiple matching (5 items) Part 3: True/False (10 items) Part 4: Multiple choice (5 items) Part 5: Multiple-choice Cloze (10 items)
Scoring method	
Scoring criteria	Each correct answer received one mark
Weighting	Equal weighting throughout the test

Contextual Parameters			PET (B1)		
Text Dimension			Part I, Questions 1-5		
Domain		I-1	① Social ② Work ③ Academic		
Discourse mode	Genre	I-2	① Public signs/notices ② Advertisements/leaflets/brochures ③ Letter/memo/email message ④ Magazine and newspaper article/report ⑤ Fiction books		
	Rhetorical task	I-3	① Exposition ② Argumentation/persuasion/evaluation ③ Historical biographical/autobiographical narrative		
Rhetorical organisation		I-4	The organisational structure of the text is: ① explicit ② ③ ④ ⑤ not explicit		
Subject specificity		I-5	Is the topic of the text of general interest or does it require subject-specific knowledge on the part of the reader? ① general ② ③ ④ ⑤ specific		
Cultural specificity		I-6	Is the topic of the text culture neutral or is it loaded with specific cultural content? ① culture neutral ② ③ ④ ⑤ culture specific		
Text abstractness		I-7	Is the text concrete or abstract? ① concrete ② ③ ④ ⑤ abstract		
Item dimension					
1	Content dimension	1-1	① Main idea ② Detail	1-2	① Fact ② Opinion
	Explicitness dimension	1-3	① from explicit information ② from implicit information		
	Did you find the information to answer the question ________?	1-4	① within a sentence ② across sentences ③ at the whole text level		
2	Content dimension	2-1	① Main idea ② Detail	2-2	① Fact ② Opinion
	Explicitness dimension	2-3	① from explicit information ② from implicit information		
	Did you find the information to answer the question ________?	2-4	① within a sentence ② across sentences ③ at the whole text level		
3	Content dimension	3-1	① Main idea ② Detail	3-2	① Fact ② Opinion
	Explicitness dimension	3-3	① from explicit information ② from implicit information		
	Did you find the information to answer the question ________?	3-4	① within a sentence ② across sentences ③ at the whole text level		
4	Content dimension	4-1	① Main idea ② Detail	4-2	① Fact ② Opinion
	Explicitness dimension	4-3	① from explicit information ② from implicit information		
	Did you find the information to answer the question ________?	4-4	① within a sentence ② across sentences ③ at the whole text level		
5	Content dimension	5-1	① Main idea ② Detail	5-2	① Fact ② Opinion
	Explicitness dimension	5-3	① from explicit information ② from implicit information		
	Did you find the information to answer the question ________?	5-4	① within a sentence ② across sentences ③ at the whole text level		

Contextual Parameters			PET (B1)		
Text Dimension			Part II, Questions 6-10		
Domain		II-1	① Social ② Work ③ Academic		
Discourse mode	Genre	II-2	① Public signs/notices ② Advertisements/leaflets/brochures ③ Letter/memo/email message ④ Magazine and newspaper article/report ⑤ Fiction books		
	Rhetorical task	II-3	① Exposition ② Argumentation/persuasion/evaluation ③ Historical biographical/autobiographical narrative		
Rhetorical organisation		II-4	The organisational structure of the text is: ① explicit ② ③ ④ ⑤ not explicit		
Subject specificity		II-5	Is the topic of the text of general interest or does it require subject-specific knowledge on the part of the reader? ① general ② ③ ④ ⑤ specific		
Cultural specificity		II-6	Is the topic of the text culture neutral or is it loaded with specific cultural content? ① culture neutral ② ③ ④ ⑤ culture specific		
Text abstractness		II-7	Is the text concrete or abstract? ① concrete ② ③ ④ ⑤ abstract		
Item dimension					
6	Content dimension	6-1	① Main idea ② Detail	6-2	① Fact ② Opinion
	Explicitness dimension	6-3	① from explicit information ② from implicit information		
	Did you find the information to answer the question ______?	6-4	① within a sentence ② across sentences ③ at the whole text level		
7	Content dimension	7-1	① Main idea ② Detail	7-2	① Fact ② Opinion
	Explicitness dimension	7-3	① from explicit information ② from implicit information		
	Did you find the information to answer the question ______?	7-4	① within a sentence ② across sentences ③ at the whole text level		
8	Content dimension	8-1	① Main idea ② Detail	8-2	① Fact ② Opinion
	Explicitness dimension	8-3	① from explicit information ② from implicit information		
	Did you find the information to answer the question ______?	8-4	① within a sentence ② across sentences ③ at the whole text level		
9	Content dimension	9-1	① Main idea ② Detail	9-2	① Fact ② Opinion
	Explicitness dimension	9-3	① from explicit information ② from implicit information		
	Did you find the information to answer the question ______?	9-4	① within a sentence ② across sentences ③ at the whole text level		
10	Content dimension	10-1	① Main idea ② Detail	10-2	① Fact ② Opinion
	Explicitness dimension	10-3	① from explicit information ② from implicit information		
	Did you find the information to answer the question ______?	10-4	① within a sentence ② across sentences ③ at the whole text level		

Contextual Parameters			**PET (B1)**				
Text Dimension			**Part III, Questions 11-20**				
Domain		III-1	① Social	② Work		③ Academic	
Discourse mode	Genre	III-2	① Public signs/notices ④ Magazine and newspaper article/report	② Advertisements/leaflets/brochures ⑤ Fiction books		③ Letter/memo/email message	
	Rhetorical task	III-3	① Exposition ③ Historical biographical/autobiographical narrative	② Argumentation/persuasion/evaluation			
Rhetorical organisation		III-4	The organisational structure of the text is: ① explicit ② ③ ④ ⑤ not explicit				
Subject specificity		III-5	Is the topic of the text of general interest or does it require subject-specific knowledge on the part of the reader? ① general ② ③ ④ ⑤ specific				
Cultural specificity		III-6	Is the topic of the text culture neutral or is it loaded with specific cultural content? ① culture neutral ② ③ ④ ⑤ culture specific				
Text abstractness		III-7	Is the text concrete or abstract? ① concrete ② ③ ④ ⑤ abstract				
Item dimension							
11	Content dimension	11-1	① Main idea	② Detail	11-2	① Fact	② Opinion
	Explicitness dimension	11-3	① from explicit information			② from implicit information	
	Did you find the information to answer the question ______?	11-4	① within a sentence			② across sentences	③ at the whole text level
12	Content dimension	12-1	① Main idea	② Detail	12-2	① Fact	② Opinion
	Explicitness dimension	12-3	① from explicit information			② from implicit information	
	Did you find the information to answer the question ______?	12-4	① within a sentence			② across sentences	③ at the whole text level
13	Content dimension	13-1	① Main idea	② Detail	13-2	① Fact	② Opinion
	Explicitness dimension	13-3	① from explicit information			② from implicit information	
	Did you find the information to answer the question ______?	13-4	① within a sentence			② across sentences	③ at the whole text level
14	Content dimension	14-1	① Main idea	② Detail	14-2	① Fact	② Opinion
	Explicitness dimension	14-3	① from explicit information			② from implicit information	
	Did you find the information to answer the question ______?	14-4	① within a sentence			② across sentences	③ at the whole text level
15	Content dimension	15-1	① Main idea	② Detail	15-2	① Fact	② Opinion
	Explicitness dimension	15-3	① from explicit information			② from implicit information	
	Did you find the information to answer the question ______?	15-4	① within a sentence			② across sentences	③ at the whole text level

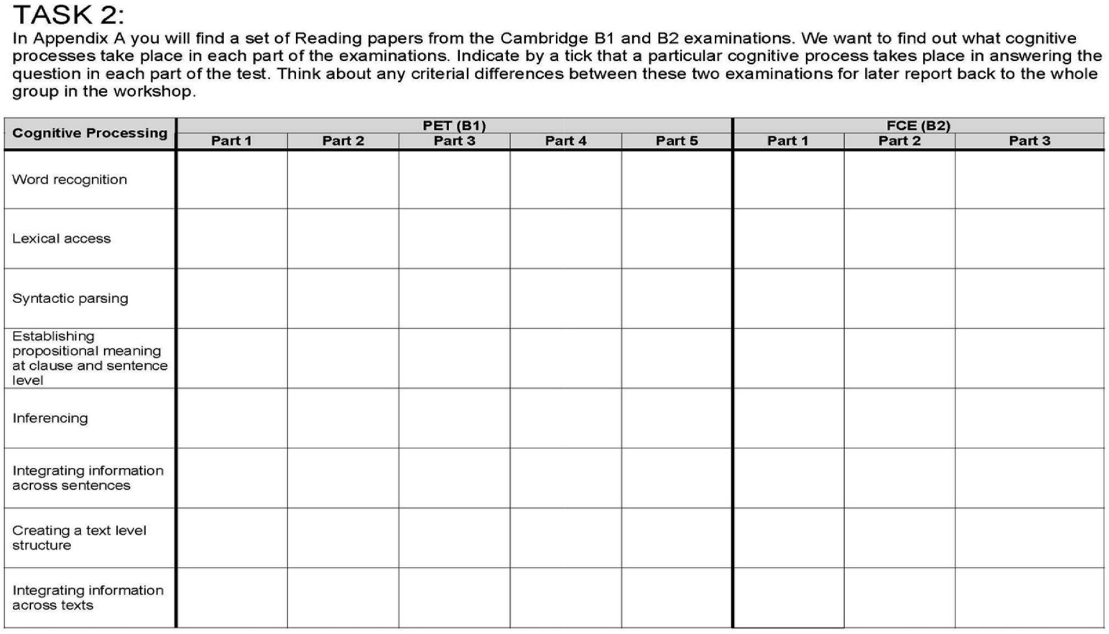
Item dimension							
16	Content dimension	16-1	① Main idea	② Detail	16-2	① Fact	② Opinion
	Explicitness dimension	16-3	① from explicit information			② from implicit information	
	Did you find the information to answer the question _______?	16-4	① within a sentence			② across sentences	③ at the whole text level
17	Content dimension	17-1	① Main idea	② Detail	17-2	① Fact	② Opinion
	Explicitness dimension	17-3	① from explicit information			② from implicit information	
	Did you find the information to answer the question _______?	17-4	① within a sentence			② across sentences	③ at the whole text level
18	Content dimension	18-1	① Main idea	② Detail	18-2	① Fact	② Opinion
	Explicitness dimension	18-3	① from explicit information			② from implicit information	
	Did you find the information to answer the question _______?	18-4	① within a sentence			② across sentences	③ at the whole text level
19	Content dimension	19-1	① Main idea	② Detail	19-2	① Fact	② Opinion
	Explicitness dimension	19-3	① from explicit information			② from implicit information	
	Did you find the information to answer the question _______?	19-4	① within a sentence			② across sentences	③ at the whole text level
20	Content dimension	20-1	① Main idea	② Detail	20-2	① Fact	② Opinion
	Explicitness dimension	20-3	① from explicit information			② from implicit information	
	Did you find the information to answer the question _______?	20-4	① within a sentence			② across sentences	③ at the whole text level

Contextual Parameters			**PET (B1)**		
Text Dimension			**Part IV, Questions 21-25**		
Domain		IV-1	① Social ② Work ③ Academic		
Discourse mode	Genre	IV-2	① Public signs/notices ② Advertisements/leaflets/brochures ③ Letter/memo/email message ④ Magazine and newspaper article/report ⑤ Fiction books		
	Rhetorical task	IV-3	① Exposition ② Argumentation/persuasion/evaluation ③ Historical biographical/autobiographical narrative		
Rhetorical organisation		IV-4	The organisational structure of the text is: ① explicit ② ③ ④ ⑤ not explicit		
Subject specificity		IV-5	Is the topic of the text of general interest or does it require subject-specific knowledge on the part of the reader? ① general ② ③ ④ ⑤ specific		
Cultural specificity		IV-6	Is the topic of the text culture neutral or is it loaded with specific cultural content? ① culture neutral ② ③ ④ ⑤ culture specific		
Text abstractness		IV-7	Is the text concrete or abstract? ① concrete ② ③ ④ ⑤ abstract		
Item dimension					
21	Content dimension	21-1	① Main idea ② Detail	21-2	① Fact ② Opinion
	Explicitness dimension	21-3	① from explicit information ② from implicit information		
	Did you find the information to answer the question _______?	21-4	① within a sentence ② across sentences ③ at the whole text level		
22	Content dimension	22-1	① Main idea ② Detail	22-2	① Fact ② Opinion
	Explicitness dimension	22-3	① from explicit information ② from implicit information		
	Did you find the information to answer the question _______?	22-4	① within a sentence ② across sentences ③ at the whole text level		
23	Content dimension	23-1	① Main idea ② Detail	23-2	① Fact ② Opinion
	Explicitness dimension	23-3	① from explicit information ② from implicit information		
	Did you find the information to answer the question _______?	23-4	① within a sentence ② across sentences ③ at the whole text level		
24	Content dimension	24-1	① Main idea ② Detail	24-2	① Fact ② Opinion
	Explicitness dimension	24-3	① from explicit information ② from implicit information		
	Did you find the information to answer the question _______?	24-4	① within a sentence ② across sentences ③ at the whole text level		
25	Content dimension	25-1	① Main idea ② Detail	25-2	① Fact ② Opinion
	Explicitness dimension	25-3	① from explicit information ② from implicit information		
	Did you find the information to answer the question _______?	25-4	① within a sentence ② across sentences ③ at the whole text level		

Contextual Parameters			PET (B1)		
Text Dimension			Part V, Questions 26-35		
Domain		V-1	① Social ② Work ③ Academic		
Discourse mode	Genre	V-2	① Public signs/notices ② Advertisements/leaflets/brochures ③ Letter/memo/email message ④ Magazine and newspaper article/report ⑤ Fiction books		
	Rhetorical task	V-3	① Exposition ② Argumentation/persuasion/evaluation ③ Historical biographical/autobiographical narrative		
Rhetorical organisation		V-4	The organisational structure of the text is: ① explicit ② ③ ④ ⑤ not explicit		
Subject specificity		V-5	Is the topic of the text of general interest or does it require subject-specific knowledge on the part of the reader? ① general ② ③ ④ ⑤ specific		
Cultural specificity		V-6	Is the topic of the text culture neutral or is it loaded with specific cultural content? ① culture neutral ② ③ ④ ⑤ culture specific		
Text abstractness		V-7	Is the text concrete or abstract? ① concrete ② ③ ④ ⑤ abstract		
Item dimension					
26	Content dimension	26-1	① Main idea ② Detail	26-2	① Fact ② Opinion
	Explicitness dimension	26-3	① from explicit information ② from implicit information		
	Did you find the information to answer the question _______?	26-4	① within a sentence ② across sentences ③ at the whole text level		
27	Content dimension	27-1	① Main idea ② Detail	27-2	① Fact ② Opinion
	Explicitness dimension	27-3	① from explicit information ② from implicit information		
	Did you find the information to answer the question _______?	27-4	① within a sentence ② across sentences ③ at the whole text level		
28	Content dimension	28-1	① Main idea ② Detail	28-2	① Fact ② Opinion
	Explicitness dimension	28-3	① from explicit information ② from implicit information		
	Did you find the information to answer the question _______?	28-4	① within a sentence ② across sentences ③ at the whole text level		
29	Content dimension	29-1	① Main idea ② Detail	29-2	① Fact ② Opinion
	Explicitness dimension	29-3	① from explicit information ② from implicit information		
	Did you find the information to answer the question _______?	29-4	① within a sentence ② across sentences ③ at the whole text level		
30	Content dimension	30-1	① Main idea ② Detail	30-2	① Fact ② Opinion
	Explicitness dimension	30-3	① from explicit information ② from implicit information		
	Did you find the information to answer the question _______?	30-4	① within a sentence ② across sentences ③ at the whole text level		

Item dimension							
31	Content dimension	31-1	① Main idea	② Detail	31-2	① Fact	② Opinion
	Explicitness dimension	31-3	① from explicit information		② from implicit information		
	Did you find the information to answer the question ________?	31-4	① within a sentence		② across sentences		③ at the whole text level
32	Content dimension	32-1	① Main idea	② Detail	32-2	① Fact	② Opinion
	Explicitness dimension	32-3	① from explicit information		② from implicit information		
	Did you find the information to answer the question ________?	32-4	① within a sentence		② across sentences		③ at the whole text level
33	Content dimension	33-1	① Main idea	② Detail	33-2	① Fact	② Opinion
	Explicitness dimension	33-3	① from explicit information		② from implicit information		
	Did you find the information to answer the question ________?	33-4	① within a sentence		② across sentences		③ at the whole text level
34	Content dimension	34-1	① Main idea	② Detail	34-2	① Fact	② Opinion
	Explicitness dimension	34-3	① from explicit information		② from implicit information		
	Did you find the information to answer the question ________?	34-4	① within a sentence		② across sentences		③ at the whole text level
35	Content dimension	35-1	① Main idea	② Detail	35-2	① Fact	② Opinion
	Explicitness dimension	35-3	① from explicit information		② from implicit information		
	Did you find the information to answer the question ________?	35-4	① within a sentence		② across sentences		③ at the whole text level

Appendix 3 Cognitive Processing Proforma



TASK 2:

In Appendix A you will find a set of Reading papers from the Cambridge B1 and B2 examinations. We want to find out what cognitive processes take place in each part of the examinations. Indicate by a tick that a particular cognitive process takes place in answering the question in each part of the test. Think about any criterial differences between these two examinations for later report back to the whole group in the workshop.

Cognitive Processing	PET (B1)					FCE (B2)		
	Part 1	Part 2	Part 3	Part 4	Part 5	Part 1	Part 2	Part 3
Word recognition								
Lexical access								
Syntactic parsing								
Establishing propositional meaning at clause and sentence level								
Inferencing								
Integrating information across sentences								
Creating a text level structure								
Integrating information across texts								

References

Alderson, J C (1993) The relationship between grammar and reading in an English for academic purposes test battery, in Douglas, D and Chapelle, C (Eds) *A New Decade Of Language Testing Research*, Alexandria: TESOL, 203–219.

Alderson, J C (2000) *Assessing Reading*, Cambridge: Cambridge University Press.

Alderson, J C (Ed) (2002) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Case Studies*, Strasbourg: Council of Europe.

Alderson, J C (2005) *Diagnosing Foreign Language Proficiency: The Interface Between Learning and Assessment*, London: Continuum.

Alderson, J C (2007) The CEFR and the need for more research, *The Modern Language Journal* 91 (4), 659–663.

Alderson J C, Figueras, N, Kuijper, H, Nold, G, Takala S and Tardieu, C (2004) *Specification For Item Development And Classification Within The CEFR: The Dutch CEFR Construct Project*, paper presented at the Workshop on Research into and with the CEFR, University of Amsterdam.

Alderson, J C, Figueras, N, Kuijper, H, Nold, G, Takala, S and Tardieu, C (2006) Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR Construct Project, *Language Assessment Quarterly* 3 (1), 3–30.

American Council on the Teaching of Foreign Languages (1999) *ACTFL Performance Guidelines for K–12 Learners*, Yonkers: American Council on the Teaching of Foreign Languages.

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (1999) *Standards For Educational and Psychological Testing*, Washington, DC: American Educational Research Association.

Anderson, R C (1974) Concretization and sentence learning, *Journal of Educational Psychology* 66 (2), 179–183.

Angoff, W H (1971) Scales, norms, and equivalent scores, in Thorndike, R L (Ed) *Educational Measurement*, 2nd edition, Washington, DC: American Council on Education, 508–600.

Association of Language Testers in Europe (2005a) *The CEFR Grids for Speaking*, available online: www.coe.int/t/dg4/linguistic/Source/ALTE%20 CEFR%20Speaking%20Grid%20INput51.pdf

Association of Language Testers in Europe (2005b) *The CEFR Grids for Writing Tasks*, available online: www.coe.int/t/dg4/linguistic/. . ./ CEFRWritingGridv3_1_presentation.doc

Baayen, R H, Piepenbrock, R and Gulikers, L (1995) *The CELEX Lexical Database*, Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Bachman, L F (1990) *Fundamental Considerations in Language Testing*, Oxford: Oxford University Press.

Bachman, L F and Palmer, A S (1996) *Language Testing in Practice*, Oxford: Oxford University Press.

Bachman, L F and Savignon, J (1986) The evaluation of communicative language proficiency: a critique of the ACTFL oral interview, *Modern Language Journal* 70 (4), 380–390.

Bachman, L F, Davidson, F, Ryan, K and Choi, I-C (1995) *An Investigation Into the Comparability of Two Tests Of English as a Foreign Language: The Cambridge TOEFL Comparability Study*, Studies in Language Testing volume 1, Cambridge: UCLES/Cambridge University Press.

Barnes, L B and Wise, S L (1991) The utility of a modified one-parameter IRT model with small samples, *Applied Measurement in Education* 4, 143–157.

Barnett, M A (1989) *More Than Meets the Eye: Foreign Language Reading, Theory and Practice*, New Jersey: Prentice Hall Refents.

Barni, M, Scaglioso, A M and Machetti, S (2010) Linking the CILS examinations to the CEFR: The A1 speaking test, in Martyniuk, W (Ed) *Aligning Tests with the CEFR: Reflections on Using The Council of Europe's Draft Manual*, Studies in Language Testing volume 33, Cambridge: UCLES/Cambridge University Press, 159–176.

Beck, I L, McKeown, M G, Sinatra, M G and Loxterman, J A (1991) Revising social studies text from a text processing perspective: Evidence of improved comprehensibility, *Reading Research Quarterly* 27, 251–276.

Berk, R A (1976) Determination of optimal cutting scores in criterion-referenced measurement, *Journal of Experimental Education* 45, 4–9.

Berman, R A (1984) Syntactic components of the foreign language reading process, in Alderson, J C and Urquhart, A H (Eds) *Reading in a Foreign Language*, London: Longman, 139–159.

Bernhardt, E B (1999) If reading is reader-based, can there be a computer-adaptive test of reading?, in Chalhoub-Deville, M (Ed) *Issues in Computer-Adaptive Testing of Reading Proficiency*, Studies in Language Testing volume 10, Cambridge: UCLES/Cambridge University Press, 1–10.

Bhatia, V K (1997) Applied genre analysis and ESP, in Miller, T (Ed) *Functional Approaches to Written Text: Classroom Applications*, Washington, DC: United States Information Agency, 134–149.

Birnbaum, A (1968) Some latent trait models and their use in inferring an examinee's ability, in Lord, F M and Novick, M R (Eds) *Statistical Theories of Mental Test Scores*, Reading: Addison-Wesley, 425–435.

Brindley, G (1986) *The Assessment of Foreign Language Proficiency: Issues and Approaches*, Adelaide: National Curriculum Resource Centre.

Brindley, G (1991) Defining language ability: The criteria for criteria, in Sarinee, A (Ed) *Current Developments in Language Testing*, available online: www.eric.ed.gov/PDFS/ED365150.pdf

Byrnes, H (2007) Perspectives, *Modern Language Journal* 91(4), 641–645.

Cambridge ESOL (2004) *Preliminary English Test Handbook for Teachers*, Cambridge: Cambridge ESOL.

Cambridge ESOL (2007) *First Certificate in English Handbook for Teachers*, Cambridge: Cambridge ESOL.

Cambridge ESOL (2008) *Top Tips for FCE Cambridge ESOL Examinations*, Cambridge: Cambridge ESOL.

Cambridge ESOL (2009) *Preliminary English Test Handbook for Teachers*, Cambridge: Cambridge ESOL.

Camilli, G (1988) Scale shrinkage and estimation of latent distribution parameters, *Journal of Educational Statistics* 13, 227–242.

Camilli, G, Yamamoto, K and Wang, M (1993) Scale shrinkage in vertical equating, *Applied Psychological Measurement* 17, 379–388.

Canale, M (1983) On some dimensions of language proficiency, in Oller, J W (Ed) *Issues in Language Testing Research*, Rowley: Newbury House Publishers, 333–342.

Canale, M and Swain, M (1980) Theoretical bases of communicative approaches to second language teaching and testing, *Applied Linguistics* 1 (1), 1–47.

Celce-Murcia, M, Dörnyei, Z and Thurrell, S (1995) Communicative competence: A pedagogically motivated model with content specifications, *Issues in Applied Linguistics* 6, 5–35.

Chall, J and Dale, E (1995) *Readability Revisited*, Cambridge: Brookline.

Cizek, G J (Ed) (2001) *Setting Performance Standards: Concepts, Methods and Perspectives*, Mahwah: Lawrence Erlbaum Associates.

Cizek, G J and Bunch, M B (2007) *Standard Setting: A Guide to Establishing Performance Standards on Tests*, Thousand Oaks: Sage Publications Inc.

Cobb, T (2010) *VocabProfile, The Compleat Lexical Tutor*, available online: www.lextutor.ca

Coltheart, M (1981) The MRC Psycholinguistic Database, *Quarterly Journal of Experimental Psychology* 33, 497–505.

Corkill, A J, Glover, J A and Bruning, R H (1988) Advance organizers: Concrete versus abstract, *Journal of Educational Research* 82, 76–81.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.

Council of Europe (2003) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment: A Manual, Preliminary Pilot Version*, Strasbourg: Council of Europe.

Council of Europe (2005) *Relating Language Examinations to the Common European Framework of Reference for Languages Learning, Teaching, Assessment (CEFR) – Reading and Listening Items and Tasks: Pilot Samples Illustrating the Common Reference Levels in English, French, German, Italian and Spanish*, Strasbourg: Council of Europe.

Council of Europe (2009) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment: A Manual*, Strasbourg: Council of Europe.

Coxhead, A J (1998) *An Academic Word List*, English Language Institute Occasional Publication 18, Wellington: Victoria University of Wellington.

Coxhead, A J (2000) A new academic word list, *TESOL Quarterly* 34 (2), 213–238.

Crossley, S A, Greenfield, J and McNamara, D S (2008) Assessing text readability using cognitively based indices, *TESOL Quarterly* 42 (3), 475–493.

Custer, M, Omar, M H and Pomplun, M (2006) Vertical scaling with the Rasch model utilizing default and tight convergence settings with WINSTEPS and BILOG-MG, *Applied Measurement in Education* 19 (2), 133–149.

Davies, A, Brown, A, Elder, C, Hill, K, Lumley, T and McNamara, R (1999) *Dictionary of Language Testing*, Studies in Language Testing volume 7, Cambridge: UCLES/Cambridge University Press.

Davis, F B (1968) Identification of subskills in reading comprehension by Maximum Likelihood Analysis, *Reading Research Quarterly* 3, 499–545.

DeMars, C (2002) Incomplete data and item parameter estimates under JMLE and MML estimation, *Applied Measurement in Education* 15, 15–32.

Dunlea, J and Matsudaira, T (2009) Investigating the relationship between the EIKEN tests and the CEFR, in Figueras, N and Noijons, J (Eds) *Standard Setting Research and its Relevance to the CEFR*, Arnhem: Cito, 103–110.

Einstein, G O, McDaniel, M A, Owen, P D and Cote, N C (1990) Encoding and recall of texts: The importance of material appropriate processing, *Journal of Memory and Language* 29, 566–581.

Embretson, S E and Wetzel, C D (1987) Component latent models for paragraph comprehension, *Applied Psychological Measurement* 11, 175–193.

Enright, M, Grabe, W, Koda, K, Mosenthal, P, Mulcahy-Ernt, P and Schedl, M (2000) *TOEFL 2000 Reading Framework: A Working Paper*, TOEFL Monograph Series 17, Princeton: Educational Testing Service.

Ferrara, S, Perie, M and Johnson, E (2002) *Matching the Judgmental Task with Standard Setting Panelist Expertise: The Item-Descriptor (ID) Matching Procedure*, Washington, DC: American Institutes for Research.

Figueras, N and Melcion, J (2002) The Common European Framework in Catalonia, in Alderson, J C (Ed) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*: *Case Studies*, Strasbourg: Council of Europe, 13–23.

Figueras, N and Noijons, J E (2009) *Linking to the CEFR Levels: Research Perspectives*, Arnhem: Cito.

Foreign Service Institute School of Language Studies (1968) *Absolute Language Proficiency Ratings*, Washington, DC: Foreign Service Institute School of Language Studies.

Fortus, R, Coriat, R and Fund, S (1998) Prediction of item difficulty in the English subtest of Israel's inter-university psychometric entrance test, in Kunnan, A J (Ed) *Validation in Language Assessment: Selected Papers From the 17th Language Research Colloquium, Long Beach*, Mahwah: Lawrence Erlbaum Associates, 61–87.

Freebody, P and Anderson, R C (1983) Effects of vocabulary difficulty, text cohesion and schema availability on reading comprehension, *Reading Research Quarterly* 18 (3), 277–94.

Freedle, R and Kostin, I (1993) *The Prediction of TOEFL Reading Comprehension Item Difficulty for Expository Prose Passages for Three Item Types: Main Idea, Inference, and Supporting Idea Items*, TOEFL Research Reports Number: RR–93–44, Princeton: Educational Testing Service.

Fry, E (1968) A readability formula that saves time, *Journal of Reading* 11 (7), 265–271.

Galloway, V B (1987) From defining to developing proficiency: A look at the decisions, in Byrnes, H and Canale, M (Eds) *Defining and Developing Proficiency Guidelines, Implementations, and Concepts*, Lincolnwood: National Textbook Company, 25–74.

Geranpayeh, A (1994) Are score comparisons across language proficiency test batteries justified? An IELTS-TOEFL comparability study, *Edinburgh Working Papers in Applied Linguistics* 5, 50–65.

Goldman, S R (1997) Learning from text: Reflections on the past and suggestions for the future, *Discourse Processes* 23, 357–398.

Goldman, S and Rakestraw, J (2000) Structural aspects of constructing meaning from text, in Kamil, M, Mosenthal, P, Pearson, P D and Barr, R (Eds) *Handbook of Reading Research*, *Volume III*, Mahwah: Lawrence Erlbaum Associates, 311–335.

Grabe, W (2000) Reading research and its implications for reading assessment, in Kunnan, A (Ed) *Fairness and Validation in Language Assessment*, Studies in Language Testing volume 9, Cambridge: UCLES/Cambridge University Press, 226–260.

Grabe, W (2009) *Reading in a Second Language: Moving from Theory to Practice*, New York: Cambridge University Press.

Graesser, A C, McNamara, D S, Louwerse, M M and Cai, Z (2004) Coh-Metrix: Analysis of text on cohesion and language, *Behavior Research Methods, Instruments, and Computers* 36, 193–202.

Green, A, Ünaldi, A and Weir, C J (2010) Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading, *Language Testing* 27 (3), 1–22.

Hanson, B A and Beguin, A A (2002) Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design, *Applied Psychological Measurement* 26 (1), 3–24.

Haynes, M and Carr, T H (1990) Writing system background and second language reading: A component skills analysis of English reading by native speaker-readers of Chinese, in Carr, T H and Levy, B A (Eds) *Reading and its Development: Component Skills Approaches*, San Diego: Academic Press, 375–421.

Henning, G (1987) *A Guide to Language Testing*, Rowley: Newbury House Publishers.

Hiple, D (1987) A progress report on the ACTFL proficiency guidelines 1982–1986, in Byrnes, H and Canale, M (Eds) *Defining and Developing Proficiency*, Lincolnwood: National Textbook Company, 5–24.

Hughes, A (1989) *Testing for Language Teachers*, Cambridge: Cambridge University Press.

Hyland, K (2000) *Disciplinary Discourses: Social Interactions in Academic Writing*, Essex: Pearson Education.

Ingram, D E (2007) *Standards in the context of teacher accreditation*, invited plenary paper to the APEC Seminar on Standards for English and other Foreign Languages in APEC Economies, Taipei, Taiwan, 3- 5 December 2007, available online: hrd.apec.org/images/7/72/47.1.pdf

Ingram, D E and Wylie, E (1979) Australian Second Language Proficiency Ratings (ASLPR), in *Adult Migrant Education Program Teachers Manual*, Canberra: Department of Immigration and Ethnic Affairs.

Jaeger, R (1989) Certification of student competence, in Linn, R L (Ed) *Educational Measurement*, 3rd edition, New York: American Council on Education Macmillan, 485–514.

Jang, E E (2009) Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment, *Language Testing* 26 (1), 31–73.

Kaftandjieva, F (2004) *Standard setting, Section B of the Reference Supplement to the Preliminary Version of the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Strasbourg: Council of Europe.

Kecker, G and Eckes, T (2010) Putting the Manual to the test: the TestDaF – CEFR linking project, in Martyniuk, W (Ed) *Aligning Tests with the CEFR: Reflections on Using The Council of Europe's Draft Manual*, Studies in Language Testing volume 33, Cambridge: UCLES/Cambridge University Press, 50–79.

Khalifa, H, ffrench, A and Salamoura, A (2010) Maintaining alignment to the CEFR: the FCE case study, in Martyniuk, W (Ed) *Aligning Tests with the CEFR: Reflections on Using The Council of Europe's Draft Manual*, Studies in Language Testing volume 33, Cambridge: UCLES/Cambridge University Press, 80–101.

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, J (2007) *A Comparison of Calibration Methods and Proficiency Estimators for Creating IRT Vertical Scales*, available online: ir.uiowa.edu/cgi/viewcontent.cgi?article=1348&context=etd

Kim, S H and Cohen, A S (1998) A comparison of linking and concurrent calibration under item response theory, *Applied Psychological Measurement* 22, 131–143.

Kolen, M J and Brennan, R L (2004) *Test Equating, Scaling, and Linking: Methods and Practices*, 2nd edition, New York: Springer-Verlag.

Landauer, T K, Foltz, P W and Laham, D (1998) Introduction to Latent Semantic Analysis, *Discourse Processes* 25, 259–284.

Language Training and Testing Center (2005) *The GEPT Intermediate Level Past Papers–3*, Taipei: Language Training and Testing Center.

Language Training and Testing Center (2008a) *GEPT Score Reports-Elementary Level*, Taipei: Language Training and Testing Center.

Language Training and Testing Center (2008b) *GEPT Score Reports-Intermediate Level*, Taipei: Language Training and Testing Center.

Language Training and Testing Center (2008c) *GEPT Score Reports-High-Intermediate Level*, Taipei: Language Training and Testing Center.

Language Training and Testing Center (2009a) *The GEPT Intermediate Level Past Paper–4*, Taipei: Language Training and Testing Center.

Language Training and Testing Center (2009b) *The GEPT High-Intermediate Level Past Paper–4*, Taipei: Language Training and Testing Center.

Language Training and Testing Center (2010) *The GEPT High-Intermediate Level Practice Paper*, Taipei: Language Training and Testing Center.

Language Training and Testing Center (2011a) *The GEPT Intermediate Level Past Paper–5*, Taipei: Language Training and Testing Center.

Language Training and Testing Center (2011b) *The GEPT High-Intermediate Level Past Paper–5*, Taipei: Language Training and Testing Center.

Lantolf, J P and Frawley, W (1985) Oral proficiency testing: A critical analysis, *The Modern Language Journal* 69, 337–345.

Lissitz, RW and Huynh, H (2003) Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability, *Practical Assessment, Research and Evaluation* 8 (10), available online: pareonline.net/getvn.asp?v=8&n=10

Little, D (2005) The Common European Framework and the European Language Portfolio: Involving learners and their judgments in the assessment process, *Language Testing* 22 (3), 321–336.

Little, D (2007) The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy, *Modern Language Journal* 91 (4), 645–655.

Little, D and Lazenby Simpson, B (2004) Using the CEF to develop an ESL curriculum for newcomer pupils in Irish primary schools, in Morrow, K (Ed)

Insights from the Common European Framework, Oxford: Oxford University Press, 91–108.

Livingston, S A (2004) *Equating Test Scores (Without IRT)*, available online: www.ets.org/Media/Research/pdf/LIVINGSTON.pdf

Loyd, B H and Hoover, H D (1980) Vertical equating using the Rasch model, *Journal of Educational Measurement* 17, 179–193.

Lumley, T J N (1993) The notion of subskills in reading comprehension tests: An EAP example, *Language Testing* 10 (3), 211–234.

Malvern, D D and Richards, B J (1997) A new measure of lexical diversity, in Ryan, A and Wray, A (Eds) *Evolving Models of Language*, Clevedon: Multilingual Matters, 58–71.

Martyniuk, W and Noijons, J (2007) *Executive Summary of Results of a Survey on the Use of the CEFR at National Level in the Council of Europe Member States*, Strasbourg: Council of Europe.

McDaniel, M A, Blischak, D and Einstein, G O (1995) Understanding the special mnemonic characteristics of fairy tales, in Weaver, C A, Mannes, S and Fletcher, C R (Eds) *Discourse Processing: Essays in Honor of Walter Kinsch*, Mahwah: Lawrence Erlbaum Associates, 157–176.

McDaniel, M A, Anderson, D C, Einstein, G O and O'Halloran, C M (1989) Modulation of environmental reinstatement effects through encoding strategies, *American Journal of Psychology* 102, 523–548.

McDaniel, M A, Einstein, G O, Dunay, P K and Cobb, R E (1986) Encoding difficulty and memory: Toward a unifying theory, *Journal of Memory and Language* 25, 645–656.

McKeown, K, Feiner, S, Robin, J, Seligmann, D and Tanenblatt, M (1992) Generating cross-references for multimedia explanation, *Proceedings of AAAI* 93, 9–16.

McNamara, D S, Louwerse, M M and Graesser, A C (2002) *Coh-Metrix: Automated Cohesion and Coherence Scores to Predict Text Readability and Facilitate Comprehension*, Institute for Intelligent Systems, University of Memphis: Memphis.

McNamara, T (1996) *Measuring Second Language Performance*, Harlow: Longman.

Messick, S (1980) The validity and the ethics of assessment, *American Psychologist* 35 (2), 1,012–1,027.

Morrow, K (2004) *Insights from the Common European Framework*, Oxford: Oxford University Press.

Munby, J L (1978) *Communicative Syllabus Design*, Cambridge: Cambridge University Press.

North, B (2000) *The Development of a Common Framework Scale of Language Proficiency*, New York: Peter Lang Publishing.

North, B (2006) *The Common European Framework of Reference: Development, theoretical and practical issues*, paper presented at A New Direction in Foreign Language Education: The Potential of the Common European Framework of Reference for Languages Symposium, Osaka University of Foreign Studies, Japan, March 2006.

Nuttall, C (1996) *Teaching Reading Skills in a Foreign Language*, Oxford: Heinemann English Language Teaching.

O'Sullivan, B (2006) *Issues in Testing Business English: The revision of the Cambridge Business English Certificates*, Studies in Language Testing volume 17, Cambridge: UCLES/Cambridge University Press.

O'Sullivan, B (2008) *City & Guilds Communicator IESOL Examination (B2) CEFR Linking Project Case Study Report, City & Guilds Research Report*, available online: cdn.cityandguilds.com/ProductDocuments/International_English/General_English/8984/Additional_documents/8984_Case_study_v1.pdf

O'Sullivan, B and Weir, C (2011) Test development and validation, in O'Sullivan, B (Ed) *Language Testing: Theories and Practices*, Oxford: Palgrave Macmillan, 13–32.

Papageorgiou, S (2007) *Relating the Trinity College London GESE and ISE Exams to the Common European Framework of Reference: Piloting of the Council of Europe Draft Manual, Final Project Report*, London: Trinity College London.

Papageorgiou, S (2010) Linking international examinations to the CEFR: the Trinity College London experience, in Martyniuk, W (Ed) *Aligning Tests with the CEFR: Reflections on Using The Council of Europe's Draft Manual*, Studies in Language Testing volume 33, Cambridge: UCLES/Cambridge University Press, 145–158.

Patz, R J (2007) *Vertical Scaling in Standards-based Educational Assessment and Accountability Systems*, Washington, DC: Council of Chief State School Offices.

Patz, R J and Yao, L (2007) Methods and models for vertical scaling, in Dorans, N J, Pommerich, M and Holland, P W (Eds) *Linking and Aligning Scores and Scales*, New York: Springer: 253–272.

Pawlikowska-Smith, G (2000) *Canadian Language Benchmarks 2000: English as a Second Language for Adults*, Ottowa: Centre for Canadian Language Benchmarks.

Pawlikowska-Smith, G (2002) *Canadian Language Benchmarks 2000: Theoretical Framework*, Ottowa: Centre for Canadian Language Benchmarks.

Perera, K (1984) *Children's Writing and Reading: Analysing Classroom Language*, Oxford: Basil Blackwell.

Perfetti, C A, Rouet, J-F and Britt, M A (1999) Toward a theory of document representation, in van Oostendorp, H and Goldman, S R (Eds) *Construction of Mental Representations During Reading*, Mahwah: Lawrence Erlbaum Associates, 99–122.

Peterson, N S, Cook, L L and Stocking, M L (1983) IRT versus conventional equating methods: A comparability study of scale stability, *Journal of Educational Statistics* 8 (2), 137–156.

Peterson, N S, Kolen, M J and Hoover, H D (1989) Scaling, norming and equating, in Linn, R L (Ed) *Educational Measurement*, 3rd edition, New York: American Council on Education/Macmillan, 221–262.

Peterson, N S, Marco, G L and Steward, E E (1982) A test of the adequacy of linear score equating models, in Holland, P W and Rubin, D E (Eds) *Test Equating*, New York: Academic Press, 71–135.

Pommerich, M, Hanson, B A, Harris, D J and Sconing, J A (2004) Issues in conducting linkages between distinct tests, *Applied Psychological Measurement* 28 (4), 247–273.

Read, J (2000) *Assessing Vocabulary*, Cambridge: Cambridge University Press.

Reckase, M D (1979) Unifactor latent trait models applied to multifactor tests: Results and implications, *Journal of Educational Statistics* 3, 207–230.

Richterich, R and Schneider, G (1992) Transparency and coherence: Why and for whom?, in North, B (Ed) *Transparency in Language Learning in Europe*, Strasbourg: Council of Europe, 43–50.

Rupp, A A, Ferne, T and Choi, H (2006) How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective, *Language Testing* 23 (4), 441–474.

Sasaki, M (2000) Effects of cultural schemata on students' test-taking processes for cloze tests: A multiple data source approach, *Language Testing* 17 (1), 85–114.

Scott, M (2009) *Wordsmith Tools 5.0*, Oxford: Oxford University Press.

Shaw, S 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.

Shiotsu, T and Weir, C J (2007) The Relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance, *Language Testing* 24 (1), 99–128.

Skaggs, G and Lissitz, R W (1985) Test equating: Relevant issues and a review of recent research, *Review of Educational Research* 56, 495–630.

Skaggs, G and Lissitz, R W (1986) IRT test equating: Relevant issues and a review of recent research, *Review of Educational Research* 56 (4), 495–529.

Stanovich, K E (2000) *Progress in Understanding Reading: Scientific Foundations and New Frontiers*, New York: Guilford.

Tannenbaum, R J and Wylie, E C (2008) *Linking English-Language Test Scores Onto the Common European Framework of Reference: An Application of Standard-Setting Methodology*, TOEFL iBT Report Number: iBT–05, Princeton: Educational Testing Service.

Taylor, L (2004) Issues of test comparability, *Research Notes* 15, 2–5.

Taylor, L and Jones, N (2006) Cambridge ESOL exams and the Common European Framework of Reference (CEFR), *Research Notes* 24, 2–5.

Thissen, D and Wainer, H (1982) Some standard errors in item response theory, *Psychometrika* 47, 397–412.

Tong, Y and Kolen, M J (2007) Comparisons of methodologies and results in vertical scaling for educational achievement tests, *Applied Measurement in Education* 20 (2), 227–253.

Trim, J (1977) *Some Possible Line of Development for an Overall Structure for a European Unit/Credit Scheme for Foreign Language Learning by Adults*, Strasbourg: Council of Europe.

Urquhart, A H (1984) The effect of rhetorical ordering on readability, in Alderson, J C and Urquhart, A H (Eds) *Reading in a Foreign Language*, London: Longman, 160–175.

Urquhart, A H and Weir, C J (1998) *Reading in a Second Language: Process, Product and Practice*, New York: Longman.

Weigle, S C (2002) *Assessing Writing*, Cambridge: Cambridge University Press.

Weir, C J (1993) *Understanding and Developing Language Tests*, London: Prentice Hall.

Weir, C J (2005a) *Language Testing and Validation: An Evidence-Based Approach*, Hampshire: Palgrave Macmillan.

Weir, C J (2005b) Limitations of the Common European Framework for developing comparable examinations and tests, *Language Testing* 22 (3), 1–20.

Weir, C J (2013) Conclusions and recommendations, in Weir, C J, Vidaković, I and Galaczi, E D, *Measured Constructs: A History of Cambridge English Language Examinations 1913–2012*, Studies in Language Testing volume 37, Cambridge: UCLES/Cambridge University Press, 420–444.

Weir, C J and Porter, D (1996) The multi-divisible or unitary nature of reading:

The language tester between Scylla and Charybdis, *Reading in a Foreign Language* 10, 1–19.

Weir, C J and Taylor, L (2011) Conclusions and recommendations, in *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press, 293-313.

Weir, C J, Hawkey, R, Green, A, Devi, S and Ünaldi, A (2009) The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university, in Thompson, P (Ed) *IELTS Research Report volume 9*, Reading: British Council/IDP Australia, 97–156.

Westhoff, G (2007) Challenges and opportunities of the CEFR for reimagining foreign language pedagogy, *The Modern Language Journal* 91 (4), 676–679.

Williams, V S L, Pommerich, M and Tissen, D (1998) A comparison of developmental scales based on Thurstone methods and item response theory, *Journal of Educational Measurement* 35, 93–107.

Wright, B D and Stone, M H (1979) *Best Test Design: Rasch Measurement*, Illinois: MESA Press.

Wu, J R W and Wu, R Y F (2010) Relating the GEPT reading comprehension tests to the CEFR, in Martyniuk, W (Ed) *Aligning Tests with the CEFR: Reflections on Using The Council of Europe's Draft Manual*, Studies in Language Testing volume 33, Cambridge: UCLES/Cambridge University Press, 204–224.

Yen, W M (1986) The choice of scale for educational measurement: An IRT perspective, *Journal of Educational Measurement* 23, 299–325.

Zieky, M and Livingston, S (1977) *Manual for Setting Standards on the Basic Skills Assessment Tests*, Princeton: Educational Testing Service.

Zimowski, M F, Muraki, E, Mislevy, R J and Bock, R D (2003) *BILOG-MG for Windows (Version 3)*, Chicago: Scientific Software International Inc.

Author index

Subject index