

David Coniam · Peter Falvey *Editors*

High- Stakes Testing

The Impact of the LPATE on English
Language Teachers in Hong Kong

 Springer

High-Stakes Testing

David Coniam · Peter Falvey
Editors

High-Stakes Testing

The Impact of the LPATE on English
Language Teachers in Hong Kong

 Springer

Editors

David Coniam
Department of Curriculum and Instruction
The Education University of Hong Kong
Hong Kong, Hong Kong

Peter Falvey
Department of Curriculum and Instruction
The Education University of Hong Kong
Hong Kong, Hong Kong

ISBN 978-981-10-6357-2 ISBN 978-981-10-6358-9 (eBook)
<https://doi.org/10.1007/978-981-10-6358-9>

Library of Congress Control Number: 2018950208

© Springer Nature Singapore Pte Ltd. 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd. The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Foreword

The role of English as a global lingua franca has developed progressively over several decades. English is associated with upward financial and social mobility, and parents everywhere are keen for their children to develop the most effective English language skills. Through the 1990s and to the present day, we have seen a massive increase all over the world in the number of pupils learning English in schools. Where the learning and teaching of English was once the preserve of the wealthy middle classes and frequently took place in private language schools, it has steadily migrated to the public sector. From being just another foreign language, English has become a fundamental part of the core curriculum both in primary and secondary education in most countries.

This rapid growth in the learning and teaching of English and its general use around the world has led to a great shortage of qualified English language teachers. Increasingly, governments have devoted significant resources to the training of English language teachers and the improvement of standards of English amongst these teachers. With an estimated 12 million teachers of English working in schools around the world, it would be fair to say that many of these teachers are challenged in terms of their both professional training and the level of English they command. In 1996, for example, less than 20% of English language teachers in Hong Kong possessed a relevant degree and teaching qualification. We see many examples which suggest that the rapid growth in the demand for English has led to some very unrealistic expectations from ministries of education and a range of other stakeholders in relation to the standards of English they expect school pupils to achieve. It is not uncommon, for example, to see the Common European Framework of Reference (CEFR) level B2 targeted as a suitable exit level from secondary school. In the survey of language competences carried out by the European Commission between 2008 and 2012 (SurveyLang), it was noted, for example, that 57% of the pupils aged 16–18 in Sweden achieved the B2 level, whereas this percentage came down to 5% in France. While we might applaud the standards of English in Sweden, we should also note that 43% of the cohort remained at B1 or below. So even in a country like Sweden, with a long tradition of high standards of English language teaching, a significant proportion of the age group does not meet targeted

standards. At the same time, expectations of what it is possible to achieve need to be managed. In Malaysia, for example, the Education Blueprint 2013 to 2025 addressed, amongst other things, levels of English in Malaysian schools and ways in which these could be improved. Benchmarking studies carried out in 2013 suggested that 20% of the pupils graduating from secondary education were achieving the B2 level. In the Blueprint, it was anticipated that by 2025, 70% of the cohort should be graduating with B2 level. It is unlikely that such an expectation could be met in such a short time frame as we can see from the example of Sweden.

In this context, *High-Stakes Testing: The Impact of the LPATE on English Language Teachers in Hong Kong* represents an important and valuable body of work in education reform, teacher development and benchmarking. Although Hong Kong may not be typical, the issues addressed in the volume are relevant beyond the Hong Kong context. As a British colony before 1997, most secondary schools purportedly taught through the medium of English although it is generally acknowledged that many of the teachers in these schools used English only in very limited ways. Post 1997, as the medium of instruction moved to Cantonese we see an increased focus by the Hong Kong Government on developing the qualifications and skills of English language teachers and a desire to improve standards as this volume clearly demonstrates.

It is disappointing that much in education reform and development is either poorly documented or not documented at all. *High-Stakes Testing: The Impact of the LPATE on English Language Teachers in Hong Kong* bucks this trend thanks in large part to its editors. The fundamental issues addressed by Coniam and Falvey and their collaborators are broadly relevant beyond Hong Kong and provide a unique historical perspective on an education reform project over a 20-year period. It is to Coniam and Falvey's credit that such a record is possible. The initial consultancy exercise in 1996 led to the development of the first version of the LPATE. This is documented in some depth, as are the extensive consultations that led to the specification of the initial LPATE and the trialling that took place to establish benchmarks as well as providing further training for teachers. The volume describes in detail how benchmarks were established but it is unfortunate, though by no means out of the ordinary, that the original terms of reference set by the Hong Kong Government for LPATE did not include any criteria by which its success could be measured. However, Coniam and Falvey's ongoing involvement in the project meant that they were able to conduct post hoc qualitative and quantitative research, which suggests that teachers believe that standards had improved over time and that LPATE had made a positive contribution to the teaching of English in Hong Kong. It is disappointing that the Hong Kong Government has consistently refused to make available to researchers significant amounts of data on pass rates and support courses available. At the same time, it is encouraging that it has allowed this volume to be published.

High-Stakes Testing: The Impact of the LPATE on English Language Teachers in Hong Kong is not intended to be a cookbook that will guide the reader on how to do benchmarking or produce tests for teachers. It is, however, a comprehensive record of a major education reform project aimed at improving the English

language skills of English teachers in Hong Kong. Many aspects are documented in detail and addressed critically. The project is documented very effectively over a period of two decades, recognises both strengths and weaknesses and is refreshingly honest in its appraisal of the weaknesses.

Cambridge, UK
December 2017

Michael Milanovic

Contents

Part I Background to High-Stakes Assessment

1	Introduction and Background to High-Stakes Assessment	3
	David Coniam and Peter Falvey	
2	Research Literature	11
	David Coniam and Peter Falvey	
3	Issues in High-Stakes Assessment	27
	David Coniam and Peter Falvey	
4	Background to the Hong Kong Education System	37
	David Coniam and Peter Falvey	
5	The Initial 1996 Consultancy Study	47
	David Coniam and Peter Falvey	
6	The English Language Benchmark Subject Committee	87
	David Coniam and Peter Falvey	
7	The Pilot Benchmark Assessment (English)	105
	David Coniam and Peter Falvey	
8	Determining Benchmarks After the PBAE	125
	David Coniam and Peter Falvey	

Part II The LPATE Enhancement Courses in Hong Kong: The Case of The Chinese University of Hong Kong

9	The LPATE Training Courses: An Initiative to Improve Teacher Language Proficiency	159
	Barley Mak and Yangyu Xiao	
10	The CUHK LPATE Training Courses: Reading and Listening . . .	179
	Barley Mak and Yangyu Xiao	

11	The CUHK LPATE Training Courses: Writing, Speaking and Classroom Language	207
	Barley Mak and Yangyu Xiao	
Part III The LPATE: A High-Stakes Assessment in Operation (2001–2007)		
12	The Operation of the LPATE (2001–2005)	239
	Alan Urmston	
13	The Revision of the LPATE	257
	Alan Urmston	
14	Maintaining Standards in the Indirectly Assessed Components of the LPATE	311
	Neil Drive	
15	Misconceptions of the LPATE in the Media: Perspectives on Educational Change	323
	Neil Drive	
Part IV How Far Have Teacher Language Standards Improved Since the Inception of the LPATE in 2001?		
16	A Quantitative Investigation of Stakeholder Perceptions	349
	David Coniam, Peter Falvey and Yangyu Xiao	
17	A Qualitative Interpretation of the Impact of the LPATE on Key Stakeholders	371
	David Coniam, Peter Falvey and Yangyu Xiao	
Part V Conclusion		
18	Concluding Comments on the Benchmarking (LPATE) Project: Strengths, Weaknesses and Constraints	399
	Peter Falvey and David Coniam	
Index	417

About the Editors

David Coniam is Chair Professor and Head of Department of the Department of Curriculum and Instruction in the Faculty of Education and Human Development at The Education University of Hong Kong, where he is a Teacher Educator, working with teachers in Hong Kong primary and secondary schools. His main publication and research interests are in language assessment, language teaching methodology and computer-assisted language learning.

Peter Falvey is a Teacher Educator, formerly a Head of Department in the Faculty of Education, The University of Hong Kong. His main publication and research interests are in language assessment, first- and second-language writing methodology and text linguistics.

Abbreviations

AACD	American Association for Counselling and Development
ACTEQ	Advisory Committee on Teacher Education and Qualifications
ACTFL	American Council on Teaching of Foreign Languages
AERA	American Educational Research Association
ALTE	The Association of Language Testers in Europe
AMEG	Association for Measurement and Evaluation in Guidance
APA	American Psychological Association
ASL	Advanced Supplementary Level examination
CE	Certification of Education examination
CEELT	Cambridge Examination in English for Language Teachers
CEFR	Common European Framework of Reference
CLA	Classroom Language Assessment
CMI	Chinese as a medium of instruction
CPE	Cambridge Proficiency in English Test
CTT	Classical test theory
CUHK	The Chinese University of Hong Kong
ECR6	Education Commission Report Number 6
ED	Education Department
EDB	Education Bureau
EdUHK	The Education University of Hong Kong
EFL	English as a foreign language
ELBSC	English Language Benchmark Subject Committee
ELL	English language learner
ELSO	Credentialing teachers of English to speakers of other languages
ELT	English language teaching
EMB	Education and Manpower Bureau
EMI	English as a medium of instruction
ESL	English as a second language
ESOL	English to speakers of other languages
FCE	First Certificate in English

GMAT	Graduate Management Admission Test
GSLPA	Graduating Students' Language Proficiency Assessment
HKALE	Hong Kong Advanced Level Examinations
HKASLE	Hong Kong Advanced Supplementary Level Examination
HKBU	Hong Kong Baptist University
HKCE	Hong Kong Certificate of Education
HKDSE	Hong Kong Diploma of Secondary Education
HKEAA	Hong Kong Examinations and Assessment Authority
HKIEd	The Hong Kong Institute of Education
HKSAR	Hong Kong Special Administrative Region
HKU	The University of Hong Kong
IELTS	International English Language Testing System
IRT	Item response theory
ITA	International Teaching Assistant
LPATE	Language Proficiency Assessment for Language Teachers of English
LPTT	Language Proficiency Test for Teachers (of Italian and Japanese as a foreign language)
MFRM	Many-facet Rasch measurement
MOI	Medium of instruction
NCME	National Council on Measurement in Education
NGTQA	Non-Graduate Teacher Qualification Assessment
OPI	Oral Proficiency Interview
OUHK	Open University of Hong Kong
PBAE	Pilot Benchmark Assessment (English)
PBC	Point biserial correlation
PET	Preliminary English Test
PGCE	Postgraduate Certificate of Education
PRC	People's Republic of China
PT	Permitted teacher
PTU	Professional Teachers' Union
QTS	Qualified teacher status
RSA	Royal Society of Arts
RT	Registered teacher
SAT	Scholastic Assessment Test (previously)
SCOLAR	Standing Committee on Language Education and Research
SD	Standard deviation
SEM	Standard error of measurement
TEM-4	Test for English major students in China
TKT	Teaching Knowledge Test
TOEFL	Test of English as a Foreign Language
UCLES	University of Cambridge Local Examinations Syndicate
UE	Use of English
WPT	Written Proficiency Test

Introduction

This introduction to *High-Stakes Testing: The Impact on English Language Teachers in Hong Kong* describes how the book was conceived, initiated and developed, provides the reader with an orientation to the structure and contents of the book, and explains why the authors feel the initiative was both useful and worthwhile, both worldwide and, especially, in Southeast Asia given the implications of current initiatives to raise language teacher standards in the region.

Although the high-stakes assessment instruments described in this book were originally initiated in 1996, the work that the book describes was still ongoing as recently as 2017. This is because, as Parts I, II and V describe, although the benchmark project to assess the language proficiency of English language teachers was first conceived in 1996, it took five years to develop the initial instructions from the Education Commission and five more years to revise the original test instruments. The authors therefore decided at the outset that a detailed account of the inception, creation, development and eventual implementation of the first set of benchmark assessment instruments was both valid and valuable for the readership. In addition, references in Part I were added and updated to bring the narrative in line with the rest of the book.

In December 1995, the Hong Kong Government's Education Commission published Report Number 6 (ECR6), passing two issues to the Advisory Committee on Teacher Education and Qualifications (ACTEQ) for the latter's consideration and action. These were:

1. That minimum language proficiency standards should be met by all teachers in their chosen medium of instruction.
2. That levels of language and professional competence ('benchmark' qualifications) should be established for all language teachers.

In December 1995, the Education and Manpower Bureau (EMB) advertised in the Hong Kong press for tenders to investigate the establishment of benchmarks for teachers of English language, Putonghua and Chinese. The time frame was four months—from April to July 1996. For English, EMB proposed that benchmarks be investigated for the following purposes:

1. To establish benchmarks for primary teachers/secondary teachers/tertiary educators,
2. To establish benchmarks for language teaching purposes/for promotional purposes,
3. To establish benchmarks for teachers of subjects other than English language (i.e. teachers of such content subjects as physics, history, mathematics) who use English as the medium of instruction.

The main editors of this book—David Coniam and Peter Falvey—were appointed to carry out the consultancy and develop and pilot assessment instruments. This is reported on in Part I of the book. It should be noted that much of Part I was written some time ago—during the development phase of the LPATE from 1996 to 2001. While some of the references in the part which were then contemporaneous may now be slightly outdated, an attempt has been made to retain those that are historically important and to update others where possible. Much of the original work has, however, been retained in order to give the reader an accurate account of the inception, development and eventual implementation of the LPATE.

This introductory part introduces the reader to the purpose in writing this book which is to provide a coherent and chronological account of the design, development and implementation of the Hong Kong benchmark project (eventually named the LPATE—Language Proficiency Assessment of Teachers (English Language) from its inception in 1996 to the preparation of this book in 2018. The background to the project is outlined, along with a relevant research review on benchmarks, standards and teacher language standards and certification and an account of Hong Kong’s educational and assessment history. A short description of each of the five parts of the book is also provided.

The purpose of this book is to provide those involved in public examinations, other scholars and potential stakeholders with an account of a long-term, high-stakes assessment project from its inception, through its development, implementation and revision as well as a post hoc research investigation of its perceived impact.

Part I—Coniam and Falvey

This data-driven section spans the period from 1996 to the first administration of the benchmark assessment in 2001, when it was renamed the Language Proficiency Assessment for Teachers (English Language), or ‘LPATE’. It chronicles the inception of the benchmarks, their constructs, creation, trialling and their first major piloting.

In the first three chapters of this part, high-stakes forms of assessment are explored—both from a general and from a language-specific perspective. The setting of standards for domains other than English proficiency is also discussed.

This part of the book proceeds from the general to the specific. That is, first, high-stakes assessment in general when the use of benchmarks is discussed, e.g. medicine, law, finance and insurance and other contexts. After these first three chapters, language benchmark issues in the context of English language teachers in Hong Kong are elaborated upon, followed by descriptions of the relevant studies that were carried out in the process of setting tentative benchmarks, the creation of a syllabus, and scales and descriptors for the direct tests of speaking, writing and classroom language, and the piloting of benchmark assessment instruments.

Chapter 1 (Coniam and Falvey)

The chapter opens with an introduction to the term ‘benchmark’ and a setting of the scene for the subsequent discussion of standards setting in an international context, setting language standards for students and developing and setting language benchmarks for teachers. The chapter explores the issue of high-stakes assessment, what constitutes a benchmark and what constitutes a language benchmark in the context of high-stakes assessment. The chapter then moves to an examination of teacher certification and teacher language certification.

Chapter 2 (Coniam and Falvey)

This chapter provides a review of the research literature in the field and considers theoretical and conceptual issues in changing assessment paradigms and setting standards for high-stakes assessment purposes. The chapter examines trends in assessment and evaluation and looks at language and language teacher standards around the world.

Chapter 3 (Coniam and Falvey)

This chapter provides an examination of the issues inherent in high-stakes assessments. Issues covered include: philosophical perspectives; policy considerations which embrace washback together with the role of different stakeholders, including government, in setting and implementing benchmarks; and different methodologies for setting standards, together with ethics and transparency in the standard-setting process.

After a discussion of whether formal tests or continuous assessment are applicable in different situations, the issue of how best to raise standards is addressed. Should standards be raised only through a test or whether (as occurred in the Hong Kong situation) funds are put aside (by government) for standards to be met through enhancement programmes rather than solely by means of assessment.

Chapter 4 (Coniam and Falvey)

This chapter describes the Hong Kong education and examination systems. It provides an account of the educational system including the medium of instruction in schools and provides details of changes to the assessment system.

Chapter 5 (Coniam and Falvey)

This chapter describes the initial 1996 Hong Kong consultancy benchmark test case study. It provides a description of the initial feasibility study (1996) which investigated two proposals:

- That minimum language proficiency standards should be met by all teachers in their chosen medium of instruction.
- That levels of language and professional competence ('benchmark' qualifications) should be established for all language teachers.

A detailed account of the consultancy, its various stages and its two major objectives is described:

- To investigate what and how prototype benchmarks might be established for lower secondary teachers of English
- To investigate the kinds of test types and assessment instruments appropriate for determining prototype benchmark levels.

Chapter 6 (Coniam and Falvey)

This chapter describes the formation of the English Language Benchmark Subject Committee (1997–1998), its role and its duties. The Committee's purpose was to produce language benchmarks specifications and an assessment syllabus for promulgation to Hong Kong teachers of English language prior to a large-scale pilot exercise—the Pilot Benchmark Assessment (English). The contentious issue of a live assessment test—a classroom language benchmark—is discussed in detail.

Chapter 7 (Coniam and Falvey)

This chapter describes the Pilot Benchmark Assessment (English) exercise (PBAE) (1998–1999). The chapter covers the set-up and administration of the PBAE, a stratified random sample piloting exercise of lower secondary teachers of English. The PBAE formed the test bed for all the constructs, their benchmarks and the associated assessment instruments which were developed in the two and a half years prior to the administration of the first live LPATE test. A brief analysis of each test component is provided, as is feedback from test takers on the different test components.

Chapter 8 (Coniam and Falvey)

This chapter discusses the issue of how benchmark levels were determined after the results of the PBAE were made available to the English Language Benchmark Subject Committee (ELBSC). It examines how analytically marked tests could be calibrated with criterion-referenced tests and how 'cut' scores might be determined for them. The chapter discusses each assessment instrument and describes the steps taken to determine an overall benchmark level for the PBAE tests.

Part I ends with a substantial appendix, Appendix 8.1., which details the methodological approaches and analytical tools used in the LPATE project.

Part II—Mak and Xiao

After the consultancy was completed and the assessment instruments were developed and piloted, the first administration of the LPATE—Language Proficiency Assessment for Teachers (English Language) as it came to be known—took place

in 2001. Amidst a wave of hostile reaction from threatened teachers, especially primary school teachers of English, the government provided a large amount of funding for Hong Kong and overseas-based development and immersion programmes for teachers who wished to sign up for them in order to attain the specified LPATE Level 3. This development is chronicled in Part II by Barley Mak and Yangyu Xiao.

This part shows how the LPATE initiative was not merely a stand-alone proficiency assessment test but that the whole benchmark initiative encompassed developmental and certification programmes—funded by government grants—that were offered at local universities and higher institutes for teachers. The work by Mak and Xiao, in three chapters, is supported by relevant data, providing detail on the upgrading and enhancement LPATE courses, together with the results of teacher feedback.

The key issue in this part centres on the concept of enhancement programmes as a government-funded alternative to a high-stakes test. This part may therefore be considered a ‘how to’ guide for potential enhancement programme course providers.

Chapter 9 (Mak and Xiao)

When the LPATE was first introduced in 2000, a number of authorised training courses were provided for Hong Kong English language teachers to help them meet or exceed the English Language Proficiency Requirements (LPR). From 2000 to 2005, a range of English language courses were provided by a total of seven different tertiary institutions in Hong Kong, two tertiary institutions in Australia and the British Council in Hong Kong.

This chapter focuses on the course providers’ perspectives. The chapter first starts with an overview of the number of courses put on by the different course providers. It then focuses mainly on the perspective of one course provider—The Chinese University of Hong Kong (CUHK)—from four key perspectives. The first perspective provides a general description of the courses offered at CUHK—including the nature and length of different courses provided, as well as the student intake over the years. The second perspective relates to a detailed account of the assessment methods, marking schemes and the quality assurance mechanisms for the different modules which constituted the programme. The third perspective focuses on feedback from participants taking these courses and in what ways they considered such courses to be beneficial to their development. The fourth perspective emphasises the programme coordinator’s own reflection on the process of running the LPATE training courses at CUHK.

This chapter contributes to a better understanding of enhancement courses which provided an alternative to a summative assessment such as the Hong Kong LPATE through a detailed account of the experience of a course provider.

Chapter 10 (Mak and Xiao)

Chapter 10 introduces the reading and listening modules of LPATE training courses which were provided at The Chinese University of Hong Kong. The two modules are introduced together in one chapter since both reading and listening are assessed

in an analytic manner in the LPATE. The chapter first introduces the two key aspects assessed in the LPATE reading and listening tests, i.e. cognitive abilities as well as linguistic skills and knowledge. It then illustrates how the reading and listening modules helped participants achieve expected language standards by giving examples of tasks used in the modules, supplemented with the course providers' interpretations.

The two modules exposed participants to many authentic tasks that were related to the educational context and comprehensively addressed and developed reading and listening skills, with the tasks described providing the opportunity to enhance different aspects of cognitive abilities and linguistic skills and knowledge. The two modules therefore attempted to enhance the reading and listening abilities that proficient English language teachers should have if they were to meet the required LPATE language standards. It was also envisaged that the modules would have the potential to raise participants' awareness of the skills and strategies needed for second-language learning—with the spin-off of participants on the LPATE enhancement programmes applying the strategies they had been practising in their own English language teaching.

Chapter 11 (Mak and Xiao)

Chapter 11 focuses on the writing, speaking and classroom language modules, which are three areas that are assessed by scales and descriptors in the LPATE. The scales and descriptors adopted in each LPATE paper, namely writing, speaking and classroom language assessment, are first introduced, followed by a presentation of the tasks that were used in different modules. The chapter focuses on how these tasks addressed the constructs assessed in the LPATE, thus helping participants to meet the language requirement. This chapter provides an understanding of how the writing, speaking and classroom language assessment modules contributed to teacher professional development.

The writing, speaking and classroom language modules were designed to help participants fulfil the requirements of the LPATE by addressing the constructs embedded in the respective assessment scales—as stated in the 2000 LPATE handbook, but also as understood by experienced teacher educators. The tasks provided in the three modules were closely associated with using language in classrooms and provided participants with the opportunity to practise the language used in the school and classroom context. The tasks also raised participants' awareness of their written and spoken language in the context of teaching, thus contributing to the development of their language proficiency in the school context.

Part III—Urmston and Drave

The LPATE was launched as a public examination in March 2001 for serving teachers of English in Hong Kong primary and secondary schools who had to attain the Language Proficiency Requirement (LPR) before September 2005. Chapters 12

and 13 by Urmston report on the operation of the LPATE during the crucial years from 2001 to 2005, and the revision project that was carried out once the deadline for the attainment of the LPR by serving teachers had passed. Chapters 14 and 15 by Drave continue the description of the work of the Hong Kong Examinations and Assessment Authority (HKEAA) in connection with standards setting.

A revision of the LPATE took place between 2006 and 2007 and was first administered in 2008. This is described by Alan Urmston in the first two chapters—Chapters 12 and 13—of Part III. The Hong Kong Examinations and Assessment Authority (HKEAA) played a large part in the administration and ongoing production of tests for the LPATE, and validation and other aspects of this work are described by Neil Drave in Chap. 14 and media coverage of the LPATE in Chap. 15.

Chapter 12 (Urmston)

Chapter 12 looks at the operationalisation of the LPATE from its launch in 2001 through to 2007, after which the revised LPATE (see Chapter 13) was adopted. After an initial slow start, when approximately 400 candidates took the Assessment—perhaps because teachers embraced the possibility that the Education Department (as it was then known) would not enforce the LPR—the LPATE went from strength to strength. The candidature increased steadily to over 2000 each administration, resulting in the HKEAA administering the assessment twice per year (March and September) from 2003 through to 2005. The chapter describes the technical aspects of test design and the operational complexities of running the assessment in the midst of clear opposition to it from some stakeholders. Issues discussed include the sociological and educational impact and consequences of such a high-stakes assessment.

Chapter 13 (Urmston)

A major consequence of the high-stakes nature of the LPATE was that questions were repeatedly raised about the reliability and validity of the assessment. Teachers who had been teaching English in schools for many years found themselves failing to reach the required Level 3 and were consequently aggrieved, questioning everything from the design of the tests, the standard and reliability of the marking, and the setting of the standard itself. After an initial slow start, when approximately 400 candidates took the assessment—perhaps because teachers embraced the possibility that the Education Department (as it was then known) would not enforce the LPR—the LPATE went from strength to strength. The candidature increased steadily to over 2000 each administration, resulting in the HKEAA administering the assessment twice per year (March and September) from 2003 through to 2005. The chapter describes the technical aspects of test design and the operational complexities of running the assessment in the midst of clear opposition to it from stakeholders. Issues discussed include the sociological and educational impact and consequences of such a high-stakes assessment, the reliability of the marking, and the setting of the standard itself. It became clear during these four years that there were deficiencies in the design of the assessment that needed to be addressed, but given the expediency with which the assessment had to be delivered during this period, there had been no opportunity to carry out any kind of review or revision.

To this end, in 2005, the Education Bureau commissioned a team to carry out a review and to recommend and implement revisions to the assessment. The review/revision project was to last for two years, with the first administration of the revised LPATE set to be launched in September 2007. This chapter describes the revision project in detail, through the review of the existing tests, the redesign of the test papers, trialling of the revised papers and the first administration of the revised LPATE in 2008.

Chapter 14 (Drave)

Given the high-stakes nature of the LPATE, it is important that appropriate standards are implemented and maintained. Without consistency in this regard, the various parties which make use of the test results would be unable to use them to make employment-related decisions. The prevailing standard must be applied equally to all candidates and the standard of performance expected must be very similar from year to year. This chapter explores the issue of standards within the context of teacher education and assessment in Hong Kong. It discusses how the standards were originally set for the LPATE and what procedures and practices were put in place to maintain them.

Chapter 15 (Drave)

This chapter presents research into public perceptions of the value of the LPATE as an instrument of change in English language education in Hong Kong. The chapter reviews media coverage of the LPATE from the years in which the assessment was most important (2003 to 2007) in the sense of certifying the largest number of serving English teachers. It reviews the press coverage afforded to the LPATE (including letters to the editor, with some opinion pieces), summarising the concerns of the various contributors, but also *critically analysing* the media discourse. The chapter reflects on the nature of good teaching, as well as on the relationship between the worlds of high-impact educational assessment and the popular press.

Part IV—Coniam, Falvey and Xiao

Part IV continues the narrative of the development of the LPATE by describing the results of a data-driven exercise, funded as a research grant by the University Grants Committee in 2015 whereby a survey of stakeholders was carried out in order to assess the perceptions of those stakeholders towards the creation and inception of the LPATE. A second, qualitative study was then conducted when an in-depth analysis of 24 of the original respondents to the survey was carried out to determine their perceptions of the impact of the LPATE. Part IV operates as a coda to the earlier parts because it was carried out 14 years after the inaugural administration of the LPATE and allowed the interviewees to reflect on the changes and challenges occasioned by the introduction of the LPATE.

Part IV provides details of the 2015–2017 HKSAR Government-funded research project that investigated the impact of the LPATE 14 years after its inception. The objectives of this research project were:

1. To investigate, quantitatively, perceptions of the extent to which English teachers' English language standards may have improved since the introduction of the LPATE in 2000,
2. To investigate, qualitatively, the perceptions of relevant stakeholders—experienced English language teachers, English panel chairs and school principals—of the effects of the LPATE policy.

This part synthesises both the quantitative and qualitative data arising from the project and provides suggestions for future developments/policy.

Chapter 16 (Coniam, Falvey and Xiao)

The quantitative data arising from the survey showed that there was agreement on the following perspectives: that English language standards were acceptable; that a minimum standard was necessary; that content teachers should face minimum standards; and that heads of department should achieve Level 4. Participants' attitudes towards the LPATE were affected by: level of school, ability band, teaching experience, number of times LPATE had been taken and highest scores attained. Participants were more likely to agree that the LPATE had had a positive impact and was necessary when they had taken the LPATE more than once; when they had longer teaching experience; and when they had obtained a higher mark in the Assessment.

Chapter 17 (Coniam, Falvey and Xiao)

The introduction of the LPATE brought home the message that English language teaching is a profession that requires adequate language proficiency, subject-matter knowledge and pedagogical skills.

There were variations in the views regarding whether English language teachers and heads of department need to take the LPATE. As most English language teachers have now become qualified through taking relevant degree courses, the LPATE is considered to be a good indicator but should not be a compulsory requirement in language teacher recruitment. In this regard, the LPATE was considered to be more relevant to potential heads of department.

From a retrospective perspective, the changes caused by the LPATE were generally positive. The respondents generally felt that there has been an improvement in English language standards, language knowledge and pedagogical skills of English teachers generally. One obvious change is that English teachers are now better trained—a positive response to the HKSAR Government's initiative to introduce the LPATE. The introduction of the LPATE also gave rise to some challenges: test quality, and the issue of trust and distrust.

Part V—Falvey and Coniam

The conclusion is provided in Chap. 18. The chapter, in two parts, first outlines what has been reported in the various parts of the book. In the second part, which consists of four parts, it then:

- describes the Constraints in the initiative,
- describes its Weaknesses,
- describes its Strengths and
- provides a Conclusion.

This part provides a conclusion to the book. It summarises the benchmark project as a whole and examines the weaknesses and strengths of the project, its implementation, its revision in 2005 and the findings of the research project of 2015-2017 to assess the perspectives of stakeholders as to its impact.

Chapter 18 (Falvey and Coniam)

This chapter provides the conclusion to the book from two perspectives. Part I is comparatively short and recaps the various parts that constitute the book.

Part II assesses the effectiveness of the LPATE within the context of educational reform and the specific context of Hong Kong in transition from British to Chinese control. The main findings, issues and lessons to be learned that arose throughout the first 20 years of the LPATE are discussed. The discussion and conclusion will be grouped under the four main headings of Constraints, Weaknesses, Strengths and Conclusion.

Part I

Background to High-Stakes Assessment

David Coniam and Peter Falvey

This data-driven section, consisting of eight chapters, is an introductory section to the book that introduces the reader to the purpose in writing this book which is to provide a coherent and chronological account of the design, development and implementation of the Hong Kong benchmark project (eventually named the LPATE—Language Proficiency Assessment of Teachers of English language) from its inception in 1996 to the present (2017). The background to the project is outlined, along with a relevant research review of benchmarks, standards and teacher language standards and certification and an account of Hong Kong’s educational and assessment history. Chapter 5 describes the initial 1996 Hong Kong consultancy benchmark test case study.

Chapter 6 describes the formation of the English Language Benchmark Subject Committee (1997–1998), its role and its duties. The Committee’s purpose was to produce language benchmarks specifications and an assessment syllabus for promulgation to Hong Kong teachers of English language prior to a large-scale pilot exercise—the Pilot Benchmark Assessment (English). The contentious issue of a live assessment test—a classroom language benchmark—is discussed in detail. Chapter 7 describes the Pilot Benchmark Assessment (English) exercise [PBAE] (1998–1999), while Chapter 8 discusses the issue of how benchmark levels were determined after the results of the PBAE were made available to the English Language Benchmark Subject Committee (ELBSC).

The purpose of this book is to provide those involved in public examinations, other scholars and potential stakeholders with an account of a long-term, high-stakes assessment project from its inception, through its development, implementation and revision as well as a post hoc research investigation of its perceived impact.

Chapter 1

Introduction and Background to High-Stakes Assessment



David Coniam and Peter Falvey

Abstract This chapter provides the reader with an introduction to benchmarks and, particularly, language benchmarks and teacher language benchmarks, their origins and their development in the twentieth century.

Introduction

Many of the buzzwords, current at the end of the twentieth century and the beginning of the twenty-first century, emerged as the world changed in terms of its commercial and service industries. The term *high-stakes assessment* is one of those terms that have become very familiar in educational contexts over the past thirty years. As greater attention is paid to the efficiency and quality of personnel, the results of assessment of performance, cost-benefit analysis of human resources, and accountability in all walks of life, the notion of high-stakes assessment has been extended to an increasing number of situations in varying contexts worldwide.

In this book, the nature of high-stakes assessment is discussed, with detailed reference to a particular high-stakes form of assessment, the assessment of teachers and, in particular, the language ability of English language teachers in Hong Kong. These teachers, from March 2001, had to demonstrate competence in terms of language benchmarks that had been developed for them. The book investigates major issues in high-stakes forms of assessment, illustrating them by investigating *why* and *how* pre-defined language standards (which are occasionally referred to as *benchmarks*) are set, i.e. initiated, researched, created, developed, trialled, moderated, established, implemented and evaluated.

D. Coniam · P. Falvey (✉)
Department of Curriculum and Instruction, Faculty of Education and Human Development,
The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, Hong Kong
e-mail: falvey@eduhk.hk

D. Coniam
e-mail: coniam@eduhk.hk

© Springer Nature Singapore Pte Ltd. 2018
D. Coniam and P. Falvey (eds.), *High-Stakes Testing*,
https://doi.org/10.1007/978-981-10-6358-9_1

In the first three chapters, high-stakes forms of assessment are discussed both from a general and from a language-specific perspective. The setting of standards for domains other than English proficiency will be discussed.

This section of the book will proceed from the general to the specific. That is, high-stakes assessment in which benchmarks are used, e.g. medicine, law, finance and insurance, will first be discussed from a general perspective. Then language benchmark issues will be discussed, followed by a detailed discussion, after the first three chapters of Sect. I, of the development and piloting of language benchmarks for English language teachers in Hong Kong.

What Is High-Stakes Assessment? What Is a Benchmark? What Is a Language Benchmark?

High-stakes assessment occurs whenever an assessment or battery of assessment instruments is used to make decisions about individuals which affect their lives in significant ways, e.g. entry tests for tertiary institutions and assessments of professional competence which can affect issues such as substantiation, promotion or termination of employment contracts.

The original use of the word benchmark was literally a mark on a bench. It was used centuries ago for measuring cloth and other goods to an agreed, specified set of criteria, e.g. two bolts of cloth would be twice the length, marked on the tailor's bench, of a standard warrior's arrow (also known as a bolt once cross-bows were invented—hence a bolt of cloth). It was a mark that could be observed and agreed upon, one that was set to a criterion known and accepted by all stakeholders.

At the time that the Hong Kong benchmark project was launched, Zairi and Leonard (1996, p. 22) described a 'benchmark' as being a term that first came into use in 1838, in ordinance survey: 'A bench mark is a mark whose height, relative to ordinance datum, has been determined by levelling' so that differences in height between established points, relative to a datum, could be calculated.

The modern, more figurative use of the term benchmark first came into everyday use circa 1884 and became common in the mid-twentieth century when standards were being set for industry. The Collins COBUILD dictionary (Sinclair et al., 1987, p. 121) has two definitions for benchmark, with the second definition embracing the modern meaning, 'something whose quality, quantity, or capability is known and which therefore can be used as a standard with which other things can be compared'. Tucker (1996, p. ix) stated that the term benchmark became linked with standards, still in the manufacturing industry, when benchmark standards were pre-specified by governments and other authorities. For example, companies that bought spare parts on contract from their suppliers would specify that screws, bolts, frames and varieties of steel should be at benchmark standards of length, width, depth and tolerance (e.g. strength under pressure).

Later still, benchmarks were used in discussions of standards in service industries, e.g. finance, insurance, air freight. Later still, the terms used by the service industries began to creep into descriptions of language standards. The language which had been used to describe standards in the service industries began to be applied to language programmes and the standards of performance that were expected from them, e.g. terms such as specifications, criteria, descriptors.

It may thus be appreciated that the original use of the term benchmark has retained its second definition of a set, agreed standard throughout its wider application in different industries and different environments and made available in manuals for users. Most educational systems have established standards for students' intended achievement in terms of benchmarks or competencies. These standards specify, explicitly, levels students are expected to reach and are criteria-referenced (i.e. based on specified and described criteria which students 'can do'), hence leading to a better alignment between curriculum and assessment (Cumming, 2009).

In the case of education, high-stakes assessment includes such tests as SATs (used for entry to USA tertiary education), GMAT (used for entry to MBAs), TOEFL (language assessment of speakers used for entry to USA tertiary institutions), IELTS used for entry to British and Australian universities and for other purposes such as immigration; and Cambridge English Assessment examinations that also are used worldwide for entry to tertiary institutions.

The use of the term benchmark in educational contexts has occurred in Australia, the UK, Canada (Canadian Language Benchmarks—a 12-point scale in English as a second language for adults—<http://www.cic.gc.ca/english/pdf/pub/language-benchmarks.pdf>—accessed November 2017) and Hong Kong. In Australia, it has been used for achieving targets (or competencies) in languages other than English and in English itself. In the Australian setting, the benchmarks for languages have been set for school learners. In Canada, the Canadian Language Benchmarks assess the candidates' language abilities in English and French, and provide task-based level descriptors, as well as standards for assessment and curriculum (Jezak, 2017). In Hong Kong, targets have been set for English, Chinese and Mathematics in primary and secondary schools. However, it is in the area of language teacher certification that benchmarks have been applied most recently in Hong Kong. The description and discussion of the latter, benchmarks for English language teachers, form the basis of the book.

Teacher Certification

The setting of standards for teacher certification is not a recent phenomenon. Indeed, going back 38 years, probably, the most striking event in the history of certification occurred when *Time Magazine* (1980) published an authentic letter, written by a teacher in the USA, which was infamously distinguished by its display of misspelling, syntactic errors and incoherence. Its publication raised public concern about the

standards of teachers. The letter, cited in Soled (1995), was a note from a teacher to a parent which read:

Scott is dropping in his studies he acts as if he don't Care. Scott want to pass in his assignment at all, he had a poem to learn and fell to do it.

Such examples are extremely damaging to the profession and are clearly unacceptable at any level of evaluation and reporting. In the USA, the report of the Holmes Group (1986) was published to answer the concerns of parents and professionals in education and was instrumental in bringing in teacher assessment, stating that:

one of five major goals is to create professionally relevant and intellectually defensible standards for entry into the profession of teaching.

This topic forms the basis of findings in Sect. IV of the book when outcomes of the research project it describes on the impact of the benchmarking initiative are discussed.

The American system of teacher assessment includes basic competencies plus pedagogic knowledge. In any country, because of the very nature of professional evaluation, teacher assessment remains a high-stakes, sensitive issue. Recent developments in education and teaching call for more accountability and the demonstration of professional competence. This has emerged as a result of increased pressure from parents and professional groups who are dissatisfied with the products of the education system. As far back as the mid-1990s, Soled (1995) noted that in a survey of public attitudes, 85% of the general public in the USA thought that teachers should be required to pass competency tests. Soled (1995) argued for teacher assessment to be addressed for two major reasons:

- to prevent incompetence in the classroom,
- as part of the solution for an educational system with problems in both teacher preparation and professional practice.

As a result of the 1986 Holmes report, a number of states in the USA introduced paper-and-pencil tests for their teachers, many of whom were already accredited, foreshadowing, as will be seen later in this volume, the complaints of serving English language teachers in Hong Kong that they were already accredited teachers of many years' experience. Such certification tests are now widespread throughout the USA (e.g. see <https://www.ets.org/praxis/>, accessed November 2017). Most USA schools now require public school teachers to pass a standardised test such as ETS' *Praxis* (Angrist & Guryan, 2008; Kane, Rockoff, & Staiger, 2008; <https://teach.com/how-to-become-a-teacher/teacher-certification-tests/>—accessed November 2017) which sets tests of reading, writing and mathematics.

One further source of concern in the USA arose in the mid-1980s when the lack of language competence of many International Teaching Assistants (ITAs) in tertiary institutions was exposed. As a result of that concern, measures were taken to establish English language standards for the ITAs that must be reached before ITAs were allowed to tutor or teach US undergraduates (see Chism & Warner, 1987 for a discussion at the first national conference on ITAs where issues about testing ITAs

were debated). The certification functions of tests ensure that ITAs have sufficient linguistic skills needed to meaningfully complete their duties (Bachman & Purpura, 2007).

Thirty-five years on from the *Time Magazine* article, an interesting scenario could be noted in Massachusetts, one of the last states in the USA to introduce assessment for teachers where, in 1998, 60% of 2000 prospective teachers failed the test. Whether qualification tests across the state improved teacher quality or not is, however, still a matter for debate.

Teacher Language Certification

Similar problems with teacher standards were encountered in Guam, a protectorate of the USA, where the quality of English language education was questioned when standards in Guam schools were compared unfavourably with standards in schools on the USA mainland. Tests of reading, writing, listening and speaking were created back in 1990 by a team led by Stansfield, Karl, and Kenyon (1990) in order to ensure that minimum agreed standards were reached by the teachers of English on Guam.

To guarantee the language standards of modern language teachers, the American Council on Teaching of Foreign Languages (ACTFL) established Oral Proficiency Interviews (OPIs) and Written Proficiency Tests (WPTs) in 1981. As English teacher proficiency is considered to be an essential characteristic for effective teaching, all students taking a Bachelor's degree in modern languages now need to obtain at least Advanced Low Level in both OPI and WPT before they are awarded Bachelor degrees and to have students to teach (Burke, 2015). OPI and WPT are considered to be high-stakes tests for university students taking a degree course in modern languages, as those who fail in the test cannot enter the teaching profession.

In Australia, the LPTT (Language Proficiency Test for Teachers of Italian and Japanese as a foreign language in Australia) was designed to assess Italian and Japanese students' language ability, including: the ability to explain subject-specific metalinguistic concepts, the ability to summarise, paraphrase, simplify information, and the ability to formulate questions and initiate classroom activities (Burke, 2015). All these tasks are similar to tasks teachers are expected to perform in classrooms (Elder, 2001).

As Elder and Kim (2014) state, teacher language certification tests should assess both 'general language proficiency' and 'academic proficiency'. General language proficiency refers to reading, listening, writing and speaking abilities in the target language, whereas academic proficiency refers to the ability to teach and the knowledge of subject-specific terminology.

A problem with current language proficiency tests is that these tests assess the speaking and writing skills, but do not guarantee classroom readiness and have limited functions in assessing communicative skills in classroom teaching.

In Hong Kong, from 2001 onwards, teachers have been expected to pass the language benchmark assessment—the LPATE—or satisfy the requirement

for exemption before they can enter the teaching profession (Bunton & Tsui, 2002; Coniam & Falvey, 2002). As stated on the Education Bureau (EDB) website, English teachers must have a relevant degree and relevant teacher training to get full exemption. Concerning requirements for exemption refer to: {<http://www.edb.gov.hk/en/teacher/qualification-training-development/qualification/language-proficiency-requirement/exemption.html>}

The LPATE test is regarded as a high-stakes test in that those who do not reach the benchmark standard (Level 3) are not allowed to teach English (see Drave, Chap. 15, this volume). Thus, the LPATE is a way of benchmarking teachers so that those who are unqualified are not permitted entry to the profession (Coniam & Falvey, 2002).

Summary

The setting of and adherence to standards in any industry or enterprise is important. It lends the product or service credibility and gives consumers a sense of assurance about the quality of the product or service. In subsequent chapters, the notion of due process will be discussed in relation to judgements which, overall, rely on humans to decide whether participants have met the standards previously set and agreed on.

References

- Angrist, J. D., & Guryan, J. (2008). Does teacher testing raise teacher quality? Evidence from state certification requirements. *Economics of Education Review*, 27(5), 483–503.
- Bachman, L. F., & Purpura, J. E. (2007). Language assessments: Gate-keepers or door-openers? In B. Spolsky, & F. M. Hult (Eds.), *The handbook of educational linguistics*. www.blackwellreference.com.
- Bunton, D., & Tsui, A. B. M. (2002). Setting language benchmarks: Whose benchmark? *Journal of Asia Pacific Communication*, 12(1), 63–76.
- Burke, B. M. (2015). A collaborative approach can improve world language education. *Phi Delta Kappa*, 96(7), 69–72.
- Chism, N., & Warner, S. B. (Eds.). (1987). *Employment and education of teaching assistants. Institutional responsibilities and responses. Readings from a national conference*. Ohio: Center for Teaching Excellence, The Ohio State University.
- Coniam, D., & Falvey, P. (2002). Selecting models and setting standards for teachers of English in Hong Kong. *Journal of Asia Pacific Communication*, 12(1), 13–37.
- Cumming, A. (2009). Language assessment in education: Tests, curricula and teaching. In B. Spolsky (Ed.), *Language policy and assessment. Special issue of annual review of applied linguistics* (vol. 29, pp. 90–100). <https://doi.org/10.1017/s0267190509090084>.
- Elder, K. (2001). Assessing the language proficiency of teachers: Are there any border controls? *Language Testing*, 18(2), 149–170.
- Elder, C., & Kim, S. (2014). Assessing teachers' language proficiency. In A. J. Kunnan (Ed.), *The companion to language assessment*. Wiley-Blackwell: Malden MA.
- Holmes Group. (1986). *Tomorrow's teachers*. East Lansing, MI: The Holmes Group Inc.
- Jejak, M. (Ed.). (2017). *Language is the key: The Canadian language benchmarks model*. Ottawa: University of Ottawa Press.

- Kane, T., Rockoff, E. R., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*, 27(6), 615–631.
- Sinclair, J., et al. (Eds.). (1987). *Collins COBUILD English language dictionary*. London and Glasgow: Collins.
- Soled, S. W. (1995). The role of assessment in teacher education. In S. W. Soled (Ed.), *Assessment testing and evaluation in teacher education*. Norwood, NJ: Ablex.
- Stansfield, C. W., Karl, J., & Kenyon, D. M. (1990). *The Guam educators' test of English proficiency (GETEP)*. Final Project Report, Revised. Washington, D.C.: Center for Applied Linguistics.
- Tucker, S. (1996). *Benchmarking*. Thousand Oaks, CA: Sage Publications.
- Zairi, M., & Leonard, P. (1996). *Origins of benchmarking and its meaning*. Dordrecht: Springer Science and Business Media.

David Coniam is Chair Professor and Head of Department of the Department of Curriculum and Instruction in the Faculty of Education and Human Development at The Education University of Hong Kong, where he is a Teacher Educator, working with teachers in Hong Kong primary and secondary schools. His main publication and research interests are in language assessment, language teaching methodology and computer-assisted language learning.

Peter Falvey is a Teacher Educator, formerly a Head of Department in the Faculty of Education, The University of Hong Kong. His main publication and research interests are in language assessment, first- and second-language writing methodology and text linguistics.

Chapter 2

Research Literature



David Coniam and Peter Falvey

Abstract In this chapter, high-stakes assessment, educational standards and benchmarks are first discussed. These are subsequently elaborated upon and discussed within their theoretical contexts and, particularly, within the context of language assessment. The ways in which assessment paradigms have changed in recent decades are also discussed to highlight their influences on the benchmarking project.

Changing Paradigms

In this chapter, high-stakes assessment, educational standards and the benchmark-setting phenomena are first placed in their theoretical contexts and then within the context of language assessment.

The major testing and assessment paradigm that was used in the last half of the twentieth century stressed the reliability of test items over their validity because of legitimate concerns about consistency and fairness in testing (Moss, 1994). In this paradigm, language tests tended to test segments of language (e.g. slot and gap-filling exercises and multiple-choice items) rather than discourse-based ‘chunks’ of language above the level of the sentence. The purpose of testing segments of language was to avoid testing elements of language other than the construct or skill being assessed. It was a paradigm that focused more on the act of *testing* than on the more holistic paradigm of *assessment*.

The connotation of the term *assessment* and, in particular, the term *high-stakes assessment* embraces a wider set of parameters than does the term *testing*. Advocates of the testing paradigm would tend to avoid any form of integrated testing. Those who were opposed to integrated forms of testing had legitimate concerns (Lee, 2006).

D. Coniam · P. Falvey (✉)
Department of Curriculum and Instruction, Faculty of Education and Human Development,
The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, Hong Kong
e-mail: falvey@eduhk.hk

D. Coniam
e-mail: coniam@eduhk.hk

Their concerns focused on the psychological distance between starting and ending the task, the potential areas for distraction or being misled during the earlier activities and the problems associated with grading/marking the final product if it had been ‘contaminated’ by what had occurred earlier (Lee, 2006). It is easy to sympathise with those who objected to integrated tests. Indeed, it was this basic stance with its aim of testing only that which should be tested that led to language being segmented so that the test items that were produced could be seen to test one element of language only.

Reliability-focused testers felt that unless a test is reliable, questions about its validity are not worth considering (Chapelle, 2012). Thus, questions of validity were shrugged off and demoted to secondary status. This perspective allowed for the growth of a form of testing known as *indirect testing*, a form of testing which states that although the indirect test may lack validity, it is possible to infer from the test taker’s score how well the skill or knowledge that is the focus of the test has been mastered (Hughes, 2003). Such tests, including the Educational Testing Services’ TOEFL dominated the language testing market for years because it could be proven that their tests were reliable (test–retest results were consistent) even though the earlier version of the TOEFL paper-based test (PBT) contained no form of direct testing of communication through speaking (an oral or speaking test) or writing (writing an extended piece of prose) (Chalhoub-Deville & Turner, 2000). The format of the TOEFL further evolved from a paper-based test to a computer-based test, then to an Internet-based test (iBT), in association with developments in the theories motivating test design; the TOEFL iBT pays more attention than did the TOEFL PBT to communicative competence and the ability to use language knowledge in relevant contexts (Taylor & Angelis, 2008).

The University of Cambridge, however, had persevered—since the inception of its Local Examinations Syndicate (UCLES) in the mid-1880s—with written tests and, later, interactive oral tests. In the face of twentieth-century concerns about the reliability of tests, it was believed by UCLES (now known as Cambridge Assessment) and its advocates that there was a place for direct tests which they later balanced with shorter segment tests of language through multiple-choice tests. In answer to criticisms about the lack of reliability on the written and spoken tests, UCLES worked hard to ensure that writing raters were standardised (all raters come to Cambridge to be trained in grading and are standardised) (Milanovic, 2016; Weir, 2005).

In addition, with the use of new technologies such as videos and DVDs, the training and standardisation of oral raters have become much more systematic and reliable. Cost, of course, is a major consideration. The results of multiple-choice tests can be scanned into a computer, and results processed very quickly (Bachman, 1990; Dooley, 2008). Direct tests, on the other hand, require human resources—assessors who have to be trained first. As will be seen below, in the midst of these paradigm changes, large testing organisations such as Educational Testing Services (ETS) have, over the past thirty years, begun to make available tests of spoken and written English to complement their original multiple-choice grammar, reading and listening tests.

Before discussing changes in English language testing, changing paradigms in testing and assessment practices in other places will be discussed. This is because changes in English language testing and assessment often follow innovations that have been made elsewhere (Eraut, 1994; Gipps, 1994). Eraut charts change to testing in the professional world of airline pilots, lawyers and doctors. Gipps (*ibid*) proposes a form of assessment that the title of her book *Beyond Testing: Towards a Theory of Educational Assessment* encapsulates. She advocates a holistic, constructive approach to assessment which de-emphasises the indirect tests, which deselect so many test takers, in favour of regular, formative assessment rather than summative assessment, and profiles of what students *can* do rather than scores which tell parents and employers little except what the student *cannot* do.

As parents and teachers reacted to old-fashioned methods of reporting the results of tests in a norm-referenced manner, Biggs (1996) illustrated this issue for Hong Kong in discussing how 80% of the school cohort at age 16 (equivalent to US Year 11) are deselected by public examinations. The problem with this form of testing and reporting is that only 20% of the whole cohort is deemed to have satisfied the examiners. The rest, having been deselected in a norm-referenced manner, have no means of showing prospective employers that they do, in fact, possess some academic or vocational qualities. Thus, Gipps (*ibid*) and Tang and Biggs (1996) advocated all-inclusive reporting of achievements for students, both stating that what should be reported is what students *can* achieve, rather than what they *cannot* achieve. Biggs (2012) reiterates that the use of criterion-referenced assessments better addresses and reflects whether and in what way students have achieved the learning objectives.

Trends in Assessment and Evaluation

Given the background above, two trends have emerged over the past four decades in the area of assessment and evaluation. These are criterion-referenced assessment (often linked to a task-based curriculum and assessment procedures in English language assessment) and competency-based assessment (often linked to vocational, and, increasingly, professional-based training and assessment) (Hudson, 2005).

On the issue of competency-based assessment, Brindley states:

Competency-based models of vocational education and training have in recent years dominated the educational landscape in Australia, the UK and New Zealand. They have also begun to exert a significant influence in the field of language learning. (1995, pp. 145–164)

Brindley (1995, pp. 1–2) stresses the need for a theoretical approach to assessment and discusses the necessity for test developers to begin with a clear theoretical conceptualisation of the abilities they are assessing and to ‘reality-test’ their constructs against data from the target language use situation. Brindley’s (1998) review of the issues inherent in outcome-based assessment and reporting in language learning programmes warns against the problems of assessing individual progress in language learning, especially when combining formative with summative reporting and in

matters of reliability and validity in outcome statements. He states that these problems can be alleviated by close consultation between policy-makers, administrators and practitioners. He discusses these issues in the context of school assessment and stresses the need for teacher professional development.

In Canada, Citizenship and Immigration Canada (CIC) decided to implement a project to develop language benchmarks for immigrants to Canada. Four TESL Canada Learners' conferences, held in 1994, discussed the issue, and the working document *Canadian Language Benchmarks* was produced in 1996. *Canadian Language Benchmarks* presented two sets of benchmarks—*Canadian Language Benchmarks: English as a Second Language for Adults* and *Canadian Language Benchmarks: English as a Second Language for Literacy Learners*. In 2000, the *Canadian Language Benchmarks 2000* was published. The work aimed at making the language benchmarks for Canada a practical and usable document (Pawlikowska-Smith, 2002). It elaborated the theoretical basis of the language benchmarks, providing examples of different language competence components, and demonstrating different levels of language proficiency, from basic language proficiency to full fluency (Fleming, 2015). In 2012, a revised version of *Canadian Language Benchmarks* was published, with an updated theoretical framework (see <http://www.cic.gc.ca/english/pdf/pub/language-benchmarks.pdf>). The benchmarks were validated against the Common European Framework for Language, the American Council for the Teaching of Foreign Languages (ACTFL) and the Quebec version of the benchmarks. The comparisons showed that the benchmarks were consistent with the theoretical concepts of the language frameworks as well as the key principles underlining other language frameworks. The validations also indicated that the Canadian Language Benchmarks were valid and reliable for a variety of purposes—including high-stakes ones—and in a variety of contexts—including community, workplace and academic (Hajer & Kaskens, 2012).

Other developed countries which possess an academically and professionally trained language teaching workforce (such as Australia) require teachers to undergo professional training before new forms of assessment can be exploited successfully (see Elder, 1994). In the area of assessment of teacher classroom language, Elder found significant correlations between the results of assessments of ESL teachers and subject specialists, indicating that agreement can be reached when assessing teachers' classroom performance. One of the major objectives in making the Hong Kong language benchmark Classroom Language Assessment (CLA) rating scales criterion-referenced was the desire for transparency so that teachers themselves, as well as informed laypersons, could, with training, reach similar grades when viewing videos of English teachers and assessing them on the four CLA scales. The Oral Proficiency Interview (OPI) of the American Council for the Teaching of Foreign Languages (ACTFL) is another example of a criterion-referenced certification test. The test results demonstrate to what extent teachers' levels of language proficiency are sufficient for them to perform teaching duties (Bachman & Purpura, 2007). A further factor in choosing criterion-referenced benchmarks is the positive washback effect which the benchmarks can engender. McDowell (1995) states:

..... it was felt that candidates for the ELSA would work towards establishing strategies for 'passing' the test by becoming test-wise and teachers would likewise seek ways to prepare their candidates to maximise their chances of success. This has already proved to be the case. (1995, p. 19) (our underlining)

The major effect of this changing paradigm of assessment on schooling is that criteria are specified for the various stages of a student's school life. Instead of being ranked against other students, the student is ranked against a series of 'can do' statements, competing against known, agreed, sets of criteria. These act as standards of achievement which schools can report to parents and, eventually, employers. Hudson (2005) reports further developments in criterion-referenced benchmark assessment, such as the Canadian Language Benchmarks, the Common European Framework and the Assessment of Language Performance Project (Hudson, 2005).

In cases where assessors have been initially trained and standardised against rating scales and descriptors, it is extremely important, for purposes of reliability, that whenever a new batch of assessments is to take place, further training is provided, particularly if there has been a significant time-gap between the initial training and the administration of the new batch of assessments (see Lumley & McNamara, 1995). Assessor training is important to ensure that all assessors assign grades in a consistent way, especially when a test is graded by a group of assessors (Sercu, 2004). However, in Baird et al.'s (2004) research, contradictory findings emerged to the effect that exemplar works and discussion about students' work did not contribute to more reliable marking. Baird et al. (2004) explain that in a well-developed community of assessment practice, one possible result of recent developments in explicit marking schemes does not necessarily need exemplars and discussions to produce accurate marking.

Language and Language Teacher Standards

Sykes and Wilson (1988) report on the work of the National Board for Professional Teaching Standards which investigated the implications of introducing procedures for the voluntary certification of teachers to a standard of 'advanced competence' with advanced levels of knowledge and skill. Foreign language teacher education and certification requirements have changed considerably over the past forty years. Whereas proficiency was not an issue in the era of audio-lingual methodology, with teachers supposedly being able to compensate for their lack of proficiency by taking their students to a language lab, foreign language teacher standards have recently become increasingly important, with more attention being paid to teachers' ability to use language in the classroom (Donato, 2009).

In developing countries, teacher certification is nowadays becoming increasingly important, although not as well established as teacher certification in developed countries (Elder & Kim, 2014; Fischer, 2013; Pearson, Fonseca-Greber, & Foell, 2006). For example, in Indonesia, the government started a national-wide teacher certification programme with the aim of certifying as many as 2.3 million teachers

by 2015, although such a certification process did not focus specifically on language standards (Fahmi, Maulana, & Yusuf, 2011). Within the context of Asian languages, Sadtono (1995) was an early caller for the certification of non-native speakers of ESL. His intentions were broadly similar to those investigated in the LPATE case study reported in this book—although his proposals did not involve the use of criterion-referenced assessments.

A discussion of some of the relevant research conducted with regard to benchmark or certification procedures in the context of second language teachers' language standards will now be presented on a country-by-country basis.

USA

In the USA, a variety of language standards agencies have been set up to guarantee the language standards of language teachers, such as the Interstate New Teacher Assessment and Support Consortium (INTASC), the American Council on the Teaching of Foreign Languages/National Council for Accreditation of Teacher Education (ACTFL/NCATE) and National Board of Professional Teaching Standards (NBPTS) (Donato, 2009).

As early as the 1990s, most USA states had some measure of certification for ESL instructors in place (see, e.g. Grant, 1995; Kornblum & Garschick, 1992; Thomas & Monoson, 1993). Many of the certification tests, however, appeared to focus on *subject-matter knowledge*, rather than on *language ability* per se, although Thomas and Monoson state, in relation to International Teaching Assistants (ITAs), that:

student complaints to legislators led to 20 states mandating higher educational institutions develop policy on oral English language proficiency of international teaching assistants. (ibid, p. 195)

The language proficiency of ITAs has been proven to be crucial in American classrooms, with research indicating how greater ITA language fluency leads to increased students' perceptions of clarity and credibility (Li, Mazer, & Ju, 2011).

In the USA, many states have made it mandatory for international students to be assessed on their oral language proficiency before they are allowed to teach. The commonly used assessments are Speaking Proficiency English Assessment Kit (SPEAK) test, locally developed teaching simulation tests, or the new TOEFL which has an integrated speaking performance component (Farnsworth, 2013). The TOEFL has been considered to measure the same general speaking proficiency as the locally developed teaching simulation tests, although the teaching simulation tests address more than general English language proficiency. The tests involve candidates giving both a prepared lecturer in a specific content area and a short explanation of course material. The Test of Oral Proficiency is such an example. The teaching simulation tests fall in line with initiatives to assess pedagogical content knowledge—an issue which Carlson (1990, pp. 157–163) discussed as far back as 1990:

to create items that require something other than a bit of pedagogical knowledge and a bit of content knowledge. (1990, p. 159)

The test designers thus attempted:

to create test items that require application of pedagogical knowledge to specific content areas. (1990, p. 160)

At the end of the twentieth century, the *No Child Left Behind* act in the USA brought further attention to teacher effectiveness and student achievement (Chambliss, 2012). Along with a concomitant increase in the number of immigrant students who were English language learners (ELL), teachers were expected to have the language knowledge to support ELL—attending, for example, to students' oral language development, supporting their academic language, and being sensible to cultural differences (Samson & Collins, 2012). Such a requirement further supported the stance that teachers need pedagogical content knowledge. Across different states in the USA—Massachusetts, New York and Florida for example—there are different teacher certification examinations focusing on the linguistic features of English, and the effective use of language to cater for different student needs (Samson & Collins, 2012).

In addition to assessing language teachers through examinations, across different states there are add-on ELSO (credentialing teachers of English to speakers of other languages) certification programmes. Certain add-on ELSO programmes have, however, been criticised for giving insufficient attention to prospective teachers' linguistic knowledge and abilities to teach linguistic knowledge to students (Reeves, 2010).

Language requirements have also been put in place for teachers of world languages (including a range of languages such as French, Spanish or German). The American Council on Teaching of Foreign Language (ACTFL) developed the OPI and WPT to assess what foreign language teachers need to know, with all language teachers having to meet minimum requirements before being permitted to take up teaching duties (Burke, 2013). The teacher language proficiency tests in the USA expect teachers to understand language acquisition and to create a supportive language classroom. Teachers are expected to use appropriate teacher talk to give instructions, to ask questions, to check learners' understanding and to guide discussions (Pearson et al., 2006).

In Sect. II, discussion will move to how issues such as those which Carlson cites above were dealt with in the development of assessment instruments used to assess the speaking and writing of English language teachers.

Canada

Since 1992, a number of documents related to language policy for teachers have been published, such as the Canadian Language Benchmarks 2000, Canadian Language Benchmarks 2012 and Language Instruction to New Comers to Canada (LINC) curriculum guidelines (Haque & Cray, 2007). These benchmarks guide teaching approaches, assessments, activities and assessments in classrooms.

In Canada, where each of the 13 Canadian provinces has its own policies for education, there are agreed expectations of language teachers. All language teachers are expected to meet three minimum requirements: a post-secondary degree from an accredited university, a qualification from an accredited teacher education programme and evidence that the person will uphold the standards of the teaching profession (Salvatori, 2009). Specifically, most Canadian jurisdictions issue a generic Kindergarten to Year 12 teacher certificate for English as a Second Language (ESL) or French as a Second Language (FSL) teachers, which limits the number of teachers who may be assigned to the teaching profession (Salvatori, 2009).

UK

Alderson et al. (1997) reported on a project with university students of French in the UK to examine their *subject-matter knowledge* in French. The University of Cambridge Local Examinations Syndicate/Royal Society of Arts examined the notion of benchmarks (*language ability and language awareness*) for entry to their Certificate in the Teaching of English as a Foreign Language programmes because of concerns that teachers applying for their courses leading to teacher examinations did not possess minimum levels of subject-matter knowledge/language awareness.

In the UK, since 1998, primary English teachers have been required to demonstrate their subject knowledge through a literacy test when exiting Postgraduate Certificate for Education programmes. Further, the Department of Education has made it a requirement that all in the state sector in the UK pass a literacy test organised by the UK Department of Education (see <https://getintoteaching.education.gov.uk/how-to-apply/passing-the-skills-tests>). The literacy professional skills test is divided into four sections: spelling, punctuation, grammar and comprehension.

The Teaching Knowledge Test (TKT) was introduced by Cambridge ESOL in 2005 for practising or trainee teachers who teach English to Speakers of Other Languages (ESOL). According to Cambridge ESOL, the test focuses on testing knowledge about teaching (Spratt, 2015). The test papers included matching and multiple-choice tasks on three modules:

1. Language and background to language learning and teaching
2. Lesson planning and use of resources for language teaching
3. Managing the teaching and learning resources (see <http://www.cambridgeenglish.org/teaching-english/teaching-qualifications/tkt/about-tkt/>).

The TKT does not assess candidates' teaching ability, performance in classroom situations or English language proficiency. Rather, it assesses candidates' declarative or received knowledge about English language teaching. The knowledge components assessed in the TKT included subject-matter knowledge, knowledge of the content, pedagogical content knowledge and general pedagogical knowledge (Spratt, 2015).

Australia

A report to the Queensland Board of Teacher Registration (1992) discussed the situation with regard to the teaching of languages other than English in Australia. McKay (1995, 2000) reported on the creation of *Bandscales* in Australia, where scales were trialled on Australian students and profiling attempted. The ESL *Bandscales* are proficiency scales tied to English language development, and they can be applied to different schools and education systems where English is a medium of instruction (McKay, 2000). Arising from the Australian experience, McKay and Ferguson (2000) debate if the *Bandscales* are transferable and in doing so present a set of questions to guide the setting of standards for L2 learners both in Australia and in Mainland China.

In an article reviewing the major contribution of Penny McKay to the field of TESOL [Note 1], Dooley and Moore (2009) noted that the *Bandscales* had three prominent features: first, the *Bandscales* were grounded in actual practice—which can reflect and guide real classrooms. Second, the *Bandscales* were informed by the assessment framework: the *Bandscales* were developed through an iterative process that interrogated researchers' and teachers' classroom observations related to second language acquisition research and related assessment research. Third, Penny McKay put considerable effort into extensive consultation with relevant stakeholders, to ensure the impact of the *Bandscales*.

Japan

In 2003, an action plan on English language education was introduced in Japan. The plan was introduced as a response to the critique that English language education focused too much on exam-related learning and did not meet the multiple needs associated with globalisation. In an attempt to ensure the quality of English language teachers, proposals were put forward for English language teachers to obtain a proficiency level equivalent to certain recognised local or international language tests. Among the proposed recognised tests and suggested levels were: the pretest level of the STEP test (the English certificate examination designed by Japan's Eiken Foundation the former *Society for Testing English Proficiency*); a score of 550 in the Test of English as a Foreign Language (TOEFL), or 730 or above in the Test of English as an International language (TOEIC) examination (Butler & Iino, 2005).

The EIKEN Test ('Jitsuyo Eigo Gino Kentei', or Test in Practical English Proficiency) is one of the most widely used English language testing programmes in Japan. The test is offered at seven levels: from Grade 1 (the highest level, equivalent to CEFR C1) down to Grade 5 (the lowest level, equivalent to CEFR A1). The EIKEN Test has been used to assess English language teachers' proficiency, with Grade Pre-1 (B2) set as the benchmark level for English teachers (Eiken Tests, 2017).

As of late 2016, Grade Pre-1 or higher on the EIKEN had been achieved by 30.2% of English teachers in middle school and by 57.3% of English teachers in high school. The target set by the Japanese government was that 50% of middle school and 75% of high school English teachers would achieve the Grade Pre-1 proficiency level by 2017 (see Kimura et al. 2017) for a discussion of developing classroom language assessment benchmarks for Japanese teachers of English.

China

The People's Republic of China (PRC) has long struggled to increase the competence and professional skills of its teachers of English language, with expansions in the provision of English language education creating a shortage of qualified English teachers. In 1988, only 30% of English teachers in junior secondary schools and 26% in senior secondary schools met the language requirements laid down by China's State Education Commission—a factor which greatly affected the envisioned goal of English language teaching in China (Hu, 2005). In addition to a considerable increase in the number of teacher education institutions and universities, there was also a substantial increase in the provision of in-service teaching training courses (Hu, 2005).

Despite the increase in provision, a coherent framework targetted specifically at English language teacher proficiency across diverse pre-service and in-service teacher education programmes in China (Hu, 2004) is still lacking. Whereas all English teachers in China are expected to pass the Test for English Major students TEM-4, such a test focuses on general English language proficiency rather than the specific English language proficiency of English teachers cited by Elder (1993).

In 2016, a national social science project at Beijing Foreign Language Studies University (Han & Qu, 2016) sets out to investigate potential ways of assessing English language teachers' language standards in Mainland China, despite the fact, as mentioned, that currently no established language standards exist for English teachers in China.

The Mainland China issue is massive and the size of the country both geographically and demographically does not lend itself to being tackled in the ways described below as in the Hong Kong case study—assessing classroom language for example. This is because of the sheer immensity of the task. It would be physically impossible to train up and standardise classroom language assessors for the whole of the English language teaching cadre in China.

Hong Kong

In Hong Kong, as long ago as the early 1990s, Tam (1992) drew attention to the need for quality control mechanisms for appraising the teaching labour force and proposed an evaluative mechanism. Within the context of secondary school teachers of English, Falvey (1995) discussed the overall lack of training that most secondary school teachers of English had received (of the cohort of 3700 in 1993, only 14.2% were both subject and professionally trained (see also Tsui, 1993). Falvey observed that:

.... with such a large proportion of the workforce unqualified to teach English, either by subject training or by professional training or both, teachers will, on the whole, have problems implementing any curriculum that requires, in addition to all other general educational knowledge and sound methodological practices, a requisite amount of subject knowledge and pedagogic content knowledge. (1995, p. 2)

Hamp-Lyons and Lumley (1998) described a project in which the development of instruments to measure the writing and speaking proficiency (for English) of all students graduating from the Hong Kong Polytechnic University was under way. The project identified domains to which the criteria of the assessment could be matched, identified actual levels of proficiency of final year students within these criteria and devised descriptions of the identified proficiencies to indicate to end-users what would be *normally* expected of graduates from various academic specialties. These researchers, were, in effect, determining benchmarks of language proficiency for English. The resulting Graduating Students' Language Proficiency Assessment (GSLPA) was first implemented in the 1999–2000 academic year. The test was a wholly workplace-oriented, task-based performance test, designed to specifically assess speaking and writing. Results on the GSLPA are reported on a scale from 1 (low) to 6 (high), with the symbol '+' indicating intermediate points on the scales, such as 2+, 3+. There are altogether 11 bands for the GSLPA (Qian, 2008). In 2007, the Hong Kong Polytechnic University, in collaboration with the then Hong Kong Institute of Education and Lingnan University, developed the Diagnostic English Language Tracking Assessment (DELTA) test, which identifies students' level of language proficiency as well as offering test takers a feedback report (Urmston, Raquel, & Tsang, 2013).

The English language benchmark case study described in this book illustrates how in Hong Kong, the initiative by the Education Commission (through its Report Number 6 in 1995) and the Advisory Committee on Teacher Education and Qualifications (ACTEQ) to set minimum language standards for *teachers* was timely, as other places in both Hong Kong and around the world began to undertake similar initiatives, using increasingly similar assessment methods.

As reported earlier, it was eventually confirmed that a battery of 'formal' tests would be created (i.e. the four skills of reading, writing, listening and speaking), together with a live classroom performance test of classroom language (Coniam & Falvey, 2002). After the tests had been assembled, a pilot study, the Pilot Benchmark Assessment (English) [PBAE], was then held in 1999 and administered to lower

secondary (Years 7–9) English language teachers to see how these coped with the prototype benchmark levels of language ability. The Government of the Hong Kong Special Administrative Region (2000) published the examination syllabus and specifications for the LPATE test in 2000. The first live administration of the tests was held in March 2001.

In early 2000, it was agreed that after 2006 the LPATE would be revised and only offered to new teachers. In addition, and with some attendant controversy, based on the studies that had been carried out, full exemption was offered to teachers holding both a relevant degree and a professional teaching qualification. At the same time, adequate provision (approximately US\$30 million) was provided by the HKSAR Government for in-service development courses for teachers to attain the required standards. A revised LPATE was completed in 2007. This was broadly similar to the 2000 version, with some minor amendments (see Urmston Chap. 14 in this volume).

From the beginning, the Hong Kong Examinations and Assessment Authority (HKEAA) administered the LPATE, with the exception of the Classroom Language Assessment component, which is administered by the Government's Education Bureau.

Making the Test Fit the Situation

While some evidence has been given above of countries which have some form of language ability benchmarking in place, the manner in which the assessment of teachers' language ability for English language teaching is conducted also needs to be taken into account. In this context, the test which appears to match closest the demands placed upon English language teachers was the Guam Educators' Test of English Proficiency (Stansfield, Karl, & Kenyon, 1990). The Guam test assessed the four skills of listening, speaking, reading and writing, with some of the test content based around the subject area of language teaching.

While the approach taken by Stansfield et al. (1990) related tasks to the teaching situation, there was still no examination of a teacher's language ability in the language classroom, which may reveal weaknesses not in evidence in a formal test situation. Furthermore, the work conducted by Stansfield and his colleagues occurred before the impetus of performance-based, criterion-related assessment was fully accepted with the result that the majority of the Guam tests contained limited-response items, with a considerable amount of multiple-choice.

Summary

This chapter has outlined a historical picture of research into language benchmarks which were beginning to be investigated and developed in various jurisdictions around the world in the 1990s and 2000s.

Note

1. Penny McKay (1948–2009), the Australian educator, who did much valuable work in this area died at a relatively young age, in the midst of her career. She and her work were honoured and commemorated in a number of articles: http://www.tesol.org.au/files/files/362_Penny_McKay_Article.pdf (Dooley & Moore, 2009).

References

- Alderson, J. C., Clapham, C., & Steel, D. (1997). Metalinguistic knowledge, language aptitude and language proficiency. *Language Teaching Research*, 1(2), 93–121.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F., & Purpura, J. E. (2007). Language assessments: Gate-keepers or door-openers? In B. Spolsky, & F. M. Hult (Eds.), *The handbook of educational linguistics*. http://www.blackwellreference.com/subscriber/uid=267/tocnode.html?id=g9781405154109_chunk_g978140515410933. Accessed November 2017.
- Baird, J. A., Grotorex, J., & Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education: Principles, Policy & Practice*, 11(3), 331–348.
- Biggs, J. (1996). The assessment scene in Hong Kong. In J. Biggs (Ed.), *Testing: To educate or to select*. Hong Kong: Hong Kong Educational Publishing Co.
- Biggs, J. (2012). What the student does: Teaching for enhanced learning. *Higher Education Research & Development*, 31(1), 39–55.
- Brindley, G. (Ed.). (1995). *Language assessment in action*. Sydney: NCELTR.
- Brindley, G. (1998). Outcomes-based assessment and reporting in language learning programmes: A review of the issues. *Language Testing*, 15(1), 45–85.
- Burke, B. M. (2013). Looking into a crystal ball: Is requiring high-stakes language proficiency tests really going to improve world language education? *Modern Language Journal*, 97(2), 531–534.
- Butler, Y. G., & Iino, M. (2005). Current Japanese reforms in English language education: The 2003 “action plan”. *Language Policy*, 4(1), 25–45.
- Carlson, R. E. (1990). Assessing teachers’ pedagogical content knowledge: Item development issues. *Journal of Personnel Evaluation in Education*, 4(2), 157–163.
- Chalhoub-Deville, M., & Turner, C. E. (2000). What to look for in ESL admission tests: Cambridge certificate exams, IELTS, and TOEFL. *System*, 28(4), 523–539.
- Chambless, K. S. (2012). Teachers’ oral proficiency in the target language: Research on its role in language teaching and learning. *Foreign Language Annals*, 48(4), 604–617.
- Chapelle, C. A. (2012). *Reliability in language assessment. The encyclopaedia of applied linguistics*. Chichester: Blackwell Publishing Ltd.
- Coniam, D., & Falvey, P. (2002). Selecting models and setting standards for teachers of English in Hong Kong. *Journal of Asian Pacific Communication*, 12(1), 13–37.
- Donato, R. (2009). Teacher education in the age of standards of professional practice. *Modern Language Journal*, 93(2), 267–270.
- Dooley, P. (2008). Language testing and technology: Problems of transition to a new era. *ReCALL*, 20(1), 21–34.
- Dooley, K., & Moore, L. (2009). Penny McKay 1948–2009: A leader in English language education. *TESOL in Context*, 19(2), 50–66.
- Eiken Tests. (2017). <http://steppeiken.org/overview-eiken-tests>. Accessed September 2017.

- Elder, C. (1993). How do subject specialists construe classroom language proficiency? *Language Testing*, 10(3), 235–254.
- Elder, C. (1994). Performance testing as a benchmark for LOTE teacher education. *Melbourne Papers in Language Testing*, 3(1), 1–25.
- Elder, C., & Kim, S. (2014). Assessing teachers' language proficiency. In A. J. Kunnan (Ed.), *The companion to language assessment*. Wiley-Blackwell: Malden, MA.
- Eraut, M. (1994). *Developing professional knowledge and competence*. London: The Falmer Press.
- Fahmi, M., Maulana, A., & Yusuf, A. (2011). Teacher certification in Indonesia: A confusion of means and ends. Working paper in economics and development studies. Padjadjaran University, Bandung: Center for Economics and Development Studies (CEDS).
- Falvey, P. (1995). The education of teachers of English in Hong Kong: A case for special treatment. In *Teacher Education in the Asian Region. Proceedings of ITEC '95* (pp. 107–113). Hong Kong: Department of Curriculum Studies, The University of Hong Kong.
- Farnsworth, T. (2013). An investigation into the validity of the TOEFL iBT speaking test for international teaching assistant certification. *Language Assessment Quarterly*, 10(3), 274–291.
- Fischer, G. (2013). Professional expectations and shattered dreams: A proficiency dilemma. *Modern Language Journal*, 97(2), 545–548.
- Fleming, D. (2015). Citizenship and race in second-language education. *Journal of Multilingual and Multicultural Development*, 36(1), 42–52.
- Gipps, C. (1994). *Beyond testing*. London: The Falmer Press.
- Grant, L. (1995). *Testing bilingual teachers' language proficiency: The case of Arizona*. Princeton, N.J.: Educational Testing Service.
- Hajer, A., & Kaskens, A. M. (2012). *Canadian language benchmarks 2012*. Ottawa: Centre for Canadian Language Benchmarks.
- Hamp-Lyons, L., & Lumley, T. (1998). *Expectations of exit language proficiency of University Graduates in Hong Kong*. Paper presented as part of 'language assessment in education in Hong Kong' colloquium at the 20th Annual Language Testing Research Colloquium, Monterey, CA.
- Han, B., & Qu, X. (2016). A comparative study of five ELT certificates. *Foreign Language Education*, 37(6), 42–47.
- Haque, E. V. E., & Cray, E. (2007). Constraining teachers: Adult ESL settlement language training policy and implementation. *TESOL Quarterly*, 41(3), 634–642.
- Hu, G. (2004). Building a strong contingent of secondary English-as-a-foreign-language teachers in China: Problems and polices. *International Journal of Educational Reform*, 14(4), 454–486.
- Hu, G. (2005). English language education in China: Policies, progress, and problems. *Language Policy*, 4(1), 5–24.
- Hudson, T. (2005). Trends in assessment scales and criterion-referenced language assessment. *Annual Review of Applied Linguistics*, 25, 205–227.
- Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.
- Kimura, Y., Nakata, Y., Ikeno, O., & Andrews, S. (2017). Developing classroom language assessment benchmarks for Japanese teachers of English as a foreign language. *Language Testing in Asia*, 7(3), 1–14. <https://doi.org/10.1186/s40468-017-0035-2>.
- Kornblum, H., & Garschick, E. (1992). *Directory of professional preparation programs in TESOL in the United States, 1992–1994*.
- Lee, Y. W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, 23(2), 131–166.
- Li, L., Mazer, J. P., & Ju, R. (2011). Resolving international teaching assistant language inadequacy through dialogue: Challenges and opportunities for clarity and credibility. *Communication Education*, 60(4), 461–478.
- Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12(1), 54–71.
- Milanovic, M. (2016). Personal communication October 2016.

- McDowell, C. (1995). Assessing the language proficiency of overseas-qualified teachers: The English Language Skills Assessment (ELSA). In G. Brindley (Ed.), *Language assessment in action* (pp. 11–29). Sydney: NCELTR.
- McKay, P. (1995). Developing ESL proficiency descriptions of the school context. In G. Brindley (Ed.), *Language assessment in action* (pp. 31–63). Sydney: National Centre for English Language Teaching and Research.
- McKay, P. (2000). On ESL profiles for school-age learners. *Language Testing*, 17(2), 185–214.
- McKay, P., & Ferguson, R. (2000). English language standards for schools in Australia and China. *Hong Kong Journal of Applied Linguistics*, 5(1), 108–127.
- Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, 23(2), 5–12.
- Pawlikowska-Smith, G. (2002). *Canadian language benchmarks 2000: Theoretical framework*. Centre for Canadian Language Benchmarks. Retrieved 8 August 2016 from <http://eric.ed.gov/?id=ED468319>.
- Pearson, L., Fonseca-Greber, B., & Foell, K. (2006). Advanced proficiency for foreign language teacher candidates: What can we do to help them achieve this goal? *Foreign Language Annals*, 39(3), 507–519.
- Qian, D. (2008). English language assessment in Hong Kong: A survey of practices, developments and issues. *Language Testing*, 25(1), 85–110.
- Reeves, J. (2010). Looking again at add-on ESOL certification. *TESOL Quarterly*, 44(2), 354–364.
- Sadtono, E. (1995, April 12–15). *The standardization of teacher trainees in EFL countries*. Paper presented at the 2nd International Conference on Language in Development: The Stakeholders' Perspectives, Denpasar Bali.
- Salvatori, M. (2009). A Canadian perspective on language teacher education: Challenges and opportunities. *The Modern Language Journal*, 93(2), 287–291.
- Samson, J. F., & Collins, B. A. (2012). *Preparing all teachers to meet the needs of English language learners: Applying research to policy and practice for teacher effectiveness*. Washington, D.C.: Center for American Progress.
- Sercu, Lies. (2004). Assessing intercultural competence: A framework for systematic test development in foreign language education and beyond. *Intercultural Education*, 15, 73–89.
- Spratt, M. (2015). TKT: Testing knowledge about teaching. In R. Wilson & M. Poulter (Eds.), *Assessing language teachers' professional skills and knowledge* (pp. 242–256). Cambridge: Cambridge University Press.
- Stansfield, C. W., Karl, J. & Kenyon, D. M. (1990). *The Guam educators' test of English Proficiency (GETEP)*. Final Project Report, Revised. Washington, D.C.: Center for Applied Linguistics.
- Sykes, G., & Wilson, S. M. (1988). *Professional standards for teaching: The assessment of teacher knowledge and skills*. Washington, DC: Office of Educational Research and Improvement.
- Tam, T. K. (1992). Quality control mechanisms for appraising the teaching labour force. *Education Journal*, 20(1), 17–24.
- Tang, C., & Biggs, J. (1996). How Hong Kong students cope with assessment. In D. A. Watkins & J. B. Biggs (Eds.), *The Chinese learner: Cultural, psychological and contextual influences*. CERC and ACER: Hong Kong and Victoria, Australia.
- Taylor, C. A., & Angelis, P. (2008). The evolution of the TOFEL. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 27–54). New York, NY: Routledge.
- Thomas, C. F., & Monoson, P. K. (1993). Oral English language proficiency of ITAs: Policy, implementation, and contributing factors. *Innovative Higher Education*, 17(3), 195–209.
- Tsui, A. B. M. (1993). *Report to the Hong Kong language campaign*. Hong Kong: Hong Kong Language Campaign.
- Urmston, A., Raquel, M., & Tsang, C. (2013). Diagnostic testing of Hong Kong tertiary students' English language proficiency: The development and validation of DELTA. *Hong Kong Journal of Applied Linguistics*, 14(2), 60–82.
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. New York, NY: Palgrave Macmillan.

David Coniam is Chair Professor and Head of Department of the Department of Curriculum and Instruction in the Faculty of Education and Human Development at The Education University of Hong Kong, where he is a teacher educator, working with teachers in Hong Kong primary and secondary schools. His main publication and research interests are in language assessment, language teaching methodology and computer assisted language learning.

Peter Falvey is a teacher educator, formerly a Head of Department in the Faculty of Education, The University of Hong Kong. His main publication and research interests are in language assessment, first and second language writing methodology, and text linguistics.

Chapter 3

Issues in High-Stakes Assessment



David Coniam and Peter Falvey

Abstract This chapter describes and discusses the major issues involved in high-stakes assessment and refers, where appropriate, to the language benchmark case study, which is described in the following chapters. The full taxonomy of major issues outlined below is not and need not always be present in its entirety in every set of benchmarks, including language benchmarks. However, most major issues need to be taken into account whenever agencies and assessment specialists meet to plan, create, establish and implement benchmarks either for the public or for specialist bodies.

Philosophical Perspectives

Chapter 2 discussed the changing paradigm in testing and assessment of all types (e.g. school, public examination, vocational assessment, etc.). When involvement in a high-stakes assessment procedure consists of stakeholders such as government, government agencies and assessment specialists, it is vital that government agencies be involved and well-briefed from the beginning. One reason for this is that government officials may not be familiar with changing paradigms or current assessment techniques. They have often been educated in an assessment environment far different from that prevailing at the time of a new assessment initiative. It is then necessary to determine how far the government, and its agencies can accept the methods proposed by the assessment specialists within the policy parameters in which they work.

In addition, government and its agencies, working together with specialist assessment consultants, are able to consider policy issues that are far broader and more far-reaching than the narrow focus which the assessment specialists, by themselves,

D. Coniam · P. Falvey (✉)
Department of Curriculum and Instruction, Faculty of Education & Human Development,
The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, Hong Kong
e-mail: falvey@eduhk.hk

D. Coniam
e-mail: coniam@eduhk.hk

© Springer Nature Singapore Pte Ltd. 2018
D. Coniam and P. Falvey (eds.), *High-Stakes Testing*,
https://doi.org/10.1007/978-981-10-6358-9_3

bring to the task. Often this leads to questions such as ‘How much of what is being proposed can be achieved’? and ‘How much is it all going to cost’?

Even more importantly, after an initial briefing and after agreement on the philosophical stance to be adopted, it is essential that there is ongoing dialogue between the government, its agencies, assessment specialists and other key stakeholders. This is because whatever may be proposed by the assessment specialists, implementation of a high-stakes assessment procedure which involves government and its agencies must fit the government policy of the time. A project which is developed over a number of months or years is often subject both to changes of government personnel (who always require additional orientation and briefing) and government policy. These changes can often frustrate specialist assessment experts. However, because policy supersedes whatever advisors might propose, advisors must learn to accept policy, personnel changes and the sociological context of the project.

Policy Considerations—Washback

Linked to the issue described above are policy decisions affecting the washback effect of high-stakes forms of assessment. The term *washback* is used in assessment to indicate that the creation of test types, test questions or test specifications will produce an effect which will wash back from the test developers to the test takers so that test taker behaviour is affected (for a detailed description of the term *washback* see Cheng & Curtis, 2012). A simple example is the effect on proficiency tests of the introduction of an oral component into a battery of tests that formerly did not contain one. The introduction of the oral test will have an immediate effect on the behaviour both of the test takers and on those who run courses for the test takers.

In the case study, described in Chap. 6, a number of policy decisions were made by the English Language Subject Benchmark Committee (ELSBC) that was set up to make recommendations to the HKSAR Government on language benchmarks for teachers. One of their major decisions was to deliberately and strongly recommend that the assessment of classroom language should take place in live classroom settings. Since no example of such procedures being carried out in other forms of teacher certification (apart from full-time professional courses run at universities for postgraduate diplomas and the RSA/UCLLES teacher certificates and diplomas) could be found at the time, washback considerations were a major issue—particularly in terms of the huge costs and logistic requirements required to carry out the assessments. The washback effect was, of course, instantaneous, with course providers for pre-service language teachers and in-service course developers immediately building into their language benchmark programmes components on classroom language awareness and practice.

The Role of Stakeholders

In the process of high-stakes assessment, there are normally a number of stakeholders, all with varying degrees of involvement in the process. Clearly, the participants in any form of benchmarking include those who initiate the process, those being assessed, those implementing the process, those who do the assessing and those who certify the process. It is possible and, indeed, likely that none of those assigned to these categories are involved in the process more than once, i.e., those being benchmarked are unlikely to be assessors of benchmarks, developers of benchmarks or implementers of benchmarks. Each role has separate and distinct functions. Other stakeholders may be trade union officials whose role may be to support and defend their members rather than to seek to be involved in setting standards, particularly if it is considered that some of its members may not reach the standards that have been set. Such a role is an uncomfortable one. Trade unions may wish to behave professionally but have to act as a defender of jobs, even though they recognise that not all their members are likely to reach set, agreed benchmarks. The role of government in maintaining high levels of information, education and the dissemination of arrangements for the implementation of benchmarks is crucial. The inclusion and engagement of as many stakeholders as possible in a benchmarking project are usually seen as a vital ingredient for the overall success of the benchmarking project.

Methodology for the Investigation—and Data Collection

The methodology used in the collection of data for any investigation is affected by the philosophical stance adopted by the researchers. Setting benchmarks requires the collection of evidence that can be analysed and interpreted so that, eventually, enough data is collected upon which to base the benchmarks. In order to create rich, ‘thick’, data during the investigative phase, it is important to use as many data sources as possible (Lincoln & Guba, 1985; Denzin & Lincoln, 2011) if they are to provide the data required to formulate benchmarks.

However, in order to begin to develop constructs for criterion-referenced benchmarks, the first form of data collection should be the sampling of the on-task performance of the clients being investigated. When it is not dangerous to carry out sampling of performances, e.g. when setting benchmarks for language teachers (or for any other cohort of professionals), and little or no disturbance is created by the collection of data, it is the investigators’ first priority to collect data in the clients’ workplace.

In the Hong Kong case study of English language teachers, described in Chap. 6, the observation of classroom language made it possible to collect, transcribe and analyse the data from which the constructs which underpin teacher language could be developed. Subsequently, descriptors for the four constructs that had been identified were created. In addition, interviews with and observation of teachers led to other

constructs (those required for the professional life of a teacher) being identified and, later, assessed.

Authenticity of Task

In high-stakes forms of assessment, test takers find it difficult to accept any form of assessment which is not, at first glance, relevant to the work they do either directly or indirectly. Bachman and Palmer (1996) defined ‘authenticity’ as the degree that test task characteristics correspond to those in Target Language Use situations. The resistance to some forms of high-stakes testing of teachers in the USA (see, e.g., the case of Massachusetts in the USA—Madaus, 1988) was fuelled by the perception that multiple-choice tests were not the best way to test a teacher’s knowledge, understanding and practice of educational principles. Authenticity has been one of the key issues addressed in language tests in recent years [see, e.g., discussion over the communicative language testing that emphasised real-life tasks and authenticity, and performance (Fulcher, 2000) and formats and model of delivery of listening tests (Taylor, 2012)]

As will be illustrated, test takers in the case study found the Classroom Language Assessment the most relevant form of assessment. They also eventually perceived the other performance tests, viz Speaking and Writing, to be authentic and linked to tasks that teachers of English have to perform. After detailed explanations of and experience of taking the other forms of assessment, they also felt that the reading tests were appropriate and relevant although problems with the Listening Test persisted for some time.

Ethics

The issue of *ethics* has always existed in high profile fields such as medicine (e.g. the role of fertility clinics, cloning and the use of brain cells and stem cells in creating life forms). However, the use of the term *ethics* is now being used regularly in academic life (the use of animals in experiments and the use of human ‘subjects’ now sometimes referred to as ‘data points’ in research).

As early as 1972, the National Council on Measurement in Education (NCME), the Association for Measurement and Evaluation in Guidance (AMEG), and the American Association for Counselling and Development (AACD is now known as the American Counselling Association) developed a position paper on the responsible use of tests that was intended to ensure that tests are given, and examinees are treated, fairly and wisely (AMEG, 1972). Later in the 1970s, AACD developed a statement on the responsibilities of the users of standardised tests, a document that was revised in 1989 (AACD, 1989). Ethical issues in assessment entered the research literature in 1972 (Schmeiser, 1995 refers to the decisions outlined in the

above paragraph). Researchers such as Hamp-Lyons and Lumley (2000), and Bailey and Butler (2004) also discuss issues such as participant involvement in assessment, the test taker's right to the release of results, issues of test taker privacy, test taker rights in the pretesting of forms of assessment, confidentiality, disclosure and anonymity. The use of indirect testing to make predictions about test takers in high-stakes assessments was beginning to be questioned at this time, hence the publication of the American Educational Research Association's (AERA) guidelines, the *Position Statement Concerning High-Stakes Testing in PreK-12 Education* (2000). The issue of ethics has also been addressed regarding accommodating test papers to cater for the needs of candidates in certain minority groups or with special needs, for example, visual, hearing or other physical impairments. However, careful consideration must be given to changes as such changes in test context, format and delivery may change the construct and inferences that can be made from the score (Taylor & Angelis, 2008).

One difficult area in ethics is the production of exemplars in high-stakes assessment procedures that contain performances by participants. This occurs when exemplar material is required for presentation purposes. In the production of video-recorded samples which show test takers taking the test it is ethically unfair to show test takers taking the test to others without first gaining the test takers' approval and indicating to them the audiences who will watch them taking the test. Prior permission must be obtained.

Transparency (Including the Need to Publish)

High-stakes examinations, fraught as they are with tension, can only benefit from attempts to make them transparent. If it is clear to the potential test taker what the benchmark is, what it consists of, what exemplars exist and whether they are easily publically available, what marking schemes are being used (made more transparent by the use of criterion-referenced assessment with its accompanying scales and descriptors), levels of anxiety are likely to decrease. As the UK Academy of Medical Royal Colleges put it (2015:7):

Since no single method and no single set of procedures can guarantee the defensibility of the standard, there is a duty of transparency towards all stakeholders around the various decisions and their implementations. Documenting how due process was followed allows the stakeholders to see the systematicity of the approach, and therefore forms part of the defensibility evidence for the standard. Following due process may at times result in uncomfortable outcomes, such as a 0% pass rate, or a different pass mark on different days of an examination. Transparency and clear communication about the process should help maintain both good practice and the acceptability of its outcomes to all stakeholders.

Part of the notion of transparency is the willingness of the 'paymaster'/the client to allow findings of ongoing investigations into high-stakes examinations to be published and disseminated. The more that can be added to the public domain the higher the level of transparency of the assessment being considered. The authors were grate-

ful to the HKSAR Government and the HKEAA for allowing them to publish the findings of the investigations they carried out into the validity of the LPATE.

Time Frames—Lead-in Periods

One of the major issues in language benchmarking is the issue of lead-in time in the formulation, preparation and implementation of a battery of assessment instruments. Inevitably, there will be a tension between the time frame that the client wants and the time frame that the researchers and assessment developers feel is required in order to do the job well. The test of who has won in this struggle is the amount of time deviation from/adherence to normal practice in the development of the battery.

Issues Involving the Mixing of Criterion-Referenced Assessment and Analytically Marked Tests

One assumption, accepted by test developers worldwide is that in order to create a battery of tests large enough to satisfy the demands of a high-stakes assessment mechanism, it might be necessary to develop a mixture of tests and test types. A major issue arises when the assessment procedures consist of a mixture of criterion-referenced assessment procedures and tests that are analytically marked.

The issue becomes one of how to calibrate analytically marked tests (such as tests of reading and listening) with criterion-referenced assessments. Criterion-referenced assessment enables a test taker profile to be created where the grades/standard/benchmarks which have been achieved by the test taker can be described on the certificate or assessment report form.

Traditional forms of reading and writing have been used for many years for purposes of norm-referenced assessment. In such cases, it does not matter that one test may be more difficult or less difficult than another because each time the test is administered, it is administered to a similar whole-population cohort and is used for selection or promotion purposes because it ranks the test takers. Such a process does not match the requirements of a benchmark test because a benchmark test wants only a cut score.

However, the problem of what the ‘cut’ scores should be still has to be faced. A ‘cut’ score is required for analytically marked tests in a battery of tests which also includes criterion-referenced tests. There are a number of methods that can be used but basically they come down to two major approaches. The first is the use of expert judges using either the Nedelsky method or the Angoff method. The essence of this approach is that the judges (at least 10–20 in number) make decisions about each item in the question paper and decide whether or not a borderline-pass test taker would score/pass on that item. The sum total of these scores are then added together

and divided by the number of assessors. The figure that is reached by these means becomes the 'cut' score. This issue is addressed in much greater detail by Drave in Section III of this volume.

The other major method is to choose the criterion-referenced test in the battery which best fits the benchmarks, e.g. the Classroom Language Assessment in the LPATE case study. The grades awarded on that benchmark are then used as a basis to statistically analyse the analytically marked tests using Rasch measurement techniques. The cut scores for the analytically marked tests that are produced by this method are used as benchmarks, for example reading and listening tests.

Exemptions

This is always a contentious issue when benchmarks are being set. A normal response in industry, when dealing with materials, is that materials affected by the new (or upgraded) benchmark must conform to benchmark standards from an agreed date. When personnel are affected, time is normally allowed for existing staff to be upgraded through development programmes or for new staff to be recruited. In certain cases, when it can be shown that certain categories of personnel already meet the new or upgraded benchmarks, exemptions are permitted either on a category or case-by-case basis.

Formal Tests or Continuous Assessment?

Linked to the issue of exemption is the issue of whether to use a one-off form of benchmarking assessment through a battery of assessment instruments at designated intervals or to carry out continuous assessment over time to discover whether participants eventually meet the benchmarks. There are arguments for both types of assessment. When the benchmark involves personnel, a one-off set of assessments can accomplish a great deal quickly. It can also be used diagnostically to indicate whether and in what areas staff may require assistance in order to attain the benchmarks that they have 'failed'.

Issues Pertaining to the Case Study

A considerable amount of money (US\$ 30 million) was set aside by the HKSAR Government to allow teachers to attend development and immersion courses in order to try to attain the benchmark.

Within the context of teacher language assessment, an important issue is whether language proficiency can be divorced from knowledge and awareness of language

(subject content) and the ability to use appropriate teaching materials and resources at an appropriate level for students (pedagogic content knowledge). These issues are addressed in the case study, particularly in the Writing Test (Tasks 2a where student language errors had to be corrected and 2b where student language errors had to be explained) and the Speaking Test (Task 3 where three test takers had to discuss a student composition).

Summary

Chapter 4 describes the background to the education system in Hong Kong, and Chap. 5 describes the methodological approaches used in the benchmark case study. Chaps. 6–9 trace the history of the benchmark initiative from its origins in 1995–1996 to its validation and implementation by the HKSAR Government in 2000–2001. The remainder of Section I therefore contains six chapters, as follows:

Chapter 4	Date	An overview of the Hong Kong education and examination systems
Chapter 5	1996	An account of the study's methodology and various statistical techniques and software packages used in Chaps. 6–9
Chapter 6	1996	The initial consultancy feasibility study
Chapter 7	1997–1998	Validation studies and the work of the English Language Benchmark Subject Committee
Chapter 8	1999	The Pilot Benchmark Assessment (English) test bed study, the PBAE
Chapter 9	2000	Determining benchmarks after the PBAE

References

- Academy of Medical Royal Colleges (AMRC). (2015). *Guidance for standard setting: A framework for high-stakes postgraduate competency-based examinations*. London: UK. Retrieved December 2016, from <http://www.aomrc.org.uk/publications/reports-guidance/standard-setting-framework-postgrad-exams-1015/>.
- American Association for Counselling and Development (AACD). (1989). *The responsibilities of users of standardized tests. AACD/AMECD policy statement: The RUST statement revised*. Retrieved January, 2018 from <http://aac.ncat.edu/Resources/documents/RUST2003%20v11%20Final.pdf>.
- American Educational Research Association (AERA). (2000). *Position statement concerning high-stakes testing in Pre K-12 education*. Retrieved January, 2018 from <http://www.aera.net/About-AERA/Position-Statements>.
- Association for Measurement and Evaluation in Guidance (AMEG). (1972). The responsible use of tests: A position paper of AMEG, APGA and NCME. *Measurement and Evaluation in Guidance*, 4(2), 385–388.

- Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Bailey, A. L., & Butler, F. A. (2004). Ethical considerations in the assessment of the language and content knowledge of US school-age English learners. *Language Assessment Quarterly*, 1, 2–3.
- Cheng, L., & Curtis, A. (2012). Test impact and washback: Implications for teaching and learning. In C. Coombe, B. O’Sullivan, P. Davidson, & S. Stoyloff (Eds.), *Cambridge guide to second language assessment* (pp. 89–95). Cambridge: Cambridge University Press.
- Denzin, N. K., & Lincoln, Y. (Eds.). (2011). *The SAGE handbook of qualitative research*. CA: SAGE Publications Inc.
- Fulcher, G. (2000). The “communicative” legacy in language testing. *System*, 28(4), 483–497.
- Hamp-Lyons, L., & Lumley, T. (2000). *Ethical dilemmas in language testing: What can we actually do?* Paper presented at the 22nd Annual Language Testing Research Colloquium, Vancouver, Canada.
- Lincoln, Y., & Guba, E. (Eds.). (1985). *Naturalistic inquiry*. Newbury Park, CA: SAGE Publications Inc.
- Madaus, G. F. (1988). The distortion of teaching and testing: High-stakes testing and instruction. *Peabody Journal of Education*, 65, 29–46.
- Schmeiser, C. B. (1995). *Ethics in assessment*. Greensboro NC: ERIC Clearinghouse on Counseling and Student Services.
- Taylor, L. (2012). Ethics in language assessment. In C. A. Chapelle (Ed.) *The encyclopedia of applied linguistics*. <https://doi.org/10.1002/9781405198431.wbeal0393>.
- Taylor, C. A., & Angelis, P. (2008). The evolution of the TOFEL. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 27–54). New York, NY: Routledge.

David Coniam is Chair Professor and Head of Department of the Department of Curriculum and Instruction in the Faculty of Education and Human Development at The Education University of Hong Kong, where he is a teacher educator, working with teachers in Hong Kong primary and secondary schools. His main publication and research interests are in language assessment, language teaching methodology and computer-assisted language learning.

Peter Falvey is a teacher educator, formerly a Head of Department in the Faculty of Education, the University of Hong Kong. His main publication and research interests are in language assessment, first and second language writing methodology, and text linguistics.

Chapter 4

Background to the Hong Kong Education System



David Coniam and Peter Falvey

Abstract This chapter provides the reader with an introduction to the Hong Kong education system and the development of teacher education in Hong Kong. It should be noted that descriptions of the Hong Kong education system have been provided in other articles by Coniam and Falvey (see Coniam and Falvey in *Validating technological innovation: The introduction and implementation of onscreen marking in Hong Kong*. Springer, Singapore, pp. 1–7, 2016; Coniam & Falvey, 2013, vi; Adamson and Li in *Education and society in Hong Kong and Macao: Comparative perspectives on continuity and change*, The University of Hong Kong, Hong Kong, pp. 35–60, 2004). Readers should not, therefore, be surprised to come across similar descriptions in the current chapter. The authoritative work on the Hong Kong education system pre-1841 to 1941 is Sweeting (*Education in Hong Kong pre-1841 to 1941*. Hong Kong University Press, Hong Kong, 1990; *A phoenix transformed: The reconstruction of education in post-war Hong Kong*. Oxford University Press, Hong Kong, 1993). (See also Tang and Bray in *Journal of Educational Administration* 38(5):468–485, 2000).

Overview of the Hong Kong Education and Examination Systems

Background

Hong Kong was governed by the UK for 156 years from 1841–1997, when the territory was finally handed back to Mainland China and became the Hong Kong

D. Coniam · P. Falvey (✉)
Department of Curriculum and Instruction, Faculty of Education & Human Development,
The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, Hong Kong
e-mail: falvey@eduhk.hk

D. Coniam
e-mail: coniam@eduhk.hk

© Springer Nature Singapore Pte Ltd. 2018
D. Coniam and P. Falvey (eds.), *High-Stakes Testing*,
https://doi.org/10.1007/978-981-10-6358-9_4

Special Administrative Region (HKSAR). During that period, the education system was based on the UK model.

Received opinion was that the British brought education to Hong Kong. Sweeting (1990, p. 2), however, rejects that notion by observing that well before the advent of the British, schools already existed in Hong Kong. After the British arrived in Hong Kong, education came mainly from missionaries; e.g., Italian missionaries began to provide schooling to British and Chinese young males in 1843. The push for the education of Chinese in a British system did not begin until the rise of social awareness in the Chinese community following the 1919 May Fourth Movement and the 1934 New Life Movement in China. Educating the poor did not become a priority until they accounted for the majority of the population.

Education and Examination Structure

The structure of mainstream education for many years was nine years of compulsory schooling in Hong Kong, six in primary school and three in junior secondary school. Over the past thirty years, however, few students actually received only nine years of education. Most received at least eleven years of education. The effective structure was six years of primary education, followed by five years of secondary education leading to the first public examination, the Hong Kong Certificate of Education Examination (HKCEE), and for a few, a further two years of education leading to the Hong Kong Advanced Level Examinations (HKALE) —the precursor to university education.

However, since 1997, the date of the handover of Hong Kong to Chinese sovereignty, there have been substantial changes to the state education system. For instance, the policy toward the language of instruction changed dramatically when Chinese medium education was promulgated soon after the handover. Incidentally, the government made a policy out of reality for the majority of its students by increasing the nine years of compulsory education to twelve years as of 2009. Furthermore, a decision to change the secondary school structure to six years not five or seven was a major initiative. Now secondary education in Hong Kong resembles the structure of secondary education in China, Australia, and the USA and lasts for six years. Major changes to the education system are shown in Table 4.1.

As shown above, under the New Academic Structure (NAS), the six years of secondary education lead to the *HKDSE (Hong Kong Diploma of Secondary Education)* examination (see below for a fuller description). After the HKDSE, students move on to work or to post-secondary, vocational, or tertiary courses. Because of the restructuring of the HKDSE, most tertiary courses are now of four years' duration.

There are three main groups of schools: government; subsidized (usually administered by religious organizations and charities); and private schools. Secondary schools are 'banded' (i.e., streamed) into three bands according to the academic level of students coming from the primary sector. Band 1 is the highest band.

Table 4.1 Education system

Under British rule		Since 2009—New Academic Structure	
Education system	Examination system	Education system	Examination system
Primary—6 years		Primary—6 years	
Secondary—5 years	Hong Kong Certificate of Education Examination (HKCEE)	Junior secondary—3 years	
Upper secondary—2 years	Hong Kong Advanced Supplementary Level Examination (HKASLE) Hong Kong Advanced Level Examination (HKALE)	Senior secondary—3 years	Hong Kong Diploma of Secondary Education (HKDSE)
Tertiary—3 years	Graduation	Tertiary—4 years	Graduation

Hong Kong has always been very examination oriented. However, more continuous and formative assessment has emerged in recent years including a large move to school-based assessment (see the description of the grading system for the HKDSE below).

For decades, it was common for two primary schools to share one set of buildings with separate morning and afternoon sessions. Nowadays, however, changes to the population have resulted in the majority of primary schools being whole-day schools.

In the 2016–2017 financial year, the total budgeted government expenditure on education was HK\$84 billion (approximately US\$10.8 billion), representing 17% of total government expenditure (<http://www.gov.hk/en/about/abouthk/factsheets/docs/education.pdf>, accessed November 2017).

English Language Learning in Hong Kong

Students in Hong Kong begin learning English at Primary 1 (age six), and students receive, on average, four to six hours' English language tuition a week in primary schools, and seven to nine hours' English language tuition in secondary schools (see Nunan, 2003). It is suggested by the EDB that about 17–21% of school hours should be devoted to English language education (<https://cd.edb.gov.hk/becg/english/chapter2.html>, accessed November 2017).

Panel Chairs

Panel chairs are heads of department—teachers who are appointed to coordinate, administer and, if qualified, provide academic leadership for all the teachers in a school who teach the panel chair's subject. The panel may consist of ten staff or more and, until relatively recently, included a number of teachers whose major teaching subject was not English.

Heads of department/panel chairs in Hong Kong secondary schools undertake many professional as well as administrative duties, including selecting textbooks, inspecting tests and homework, and approving locally produced materials (Benson, 2010). The discussions over the duty of English heads of department indicate that the latter should have language, curriculum, and managerial expertise to manage their department. On this basis, there evolved, from within the school sector, a trend to accept a rather higher baseline for the appointment of heads of department (LPATE Level 4) than for English language teachers (LPATE Level 3) (Coniam and Falvey, 2002; this volume, Section IV).

Teacher Education

Teacher training has been a neglected activity until recently. Until the 1990s, it was still a worldwide phenomenon that teachers were untrained (Li & Kwo, 2004). The lack of teacher training was related to a variety of factors, such as teacher training being just a small part of the educational system; the belief that any person who completed a particular level of education could teach students at lower levels; and, possibly, the schools' budgets as trained teachers are better paid (Li & Kwo, 2004). Until 1994, all primary teachers were educated at government-run teacher training colleges. They were not graduates; instead, they entered the colleges of education for three years after they left secondary school in Year 11 or for two years after they left secondary school in Year 13. After graduation, all were titled 'Certificated Teachers'.

Before the handover in Hong Kong in 1997, teaching, compared with other professions such as medicine or law, was a semi-profession (Morris, 2004). Historically, there had been no official requirement that a person should be professionally trained before entering the teaching profession. Any person wanting to be a teacher has to apply to EDB to become either a 'registered teacher' (RT) or a 'permitted teacher' (PT). To be qualified as a 'registered teacher', a person must have obtained 'qualified teacher status' (QTS) through completing a sub-degree level certificate/diploma of education, or a bachelor degree in education, or a postgraduate certificate/diploma in education (Lee, 2013).

The low requirement for teachers was associated with the types of professional training courses provided in teacher education. Before 1920, teacher education was mainly conducted at the training school or college level, for example, in St Paul's College, Central School, and Wanchai Normal School. The first four-year undergraduate course on teacher education was launched at the University of Hong Kong (HKU) in 1920. In 1965, the Chinese University of Hong Kong (CUHK) also established a School of Education, providing pre-service and in-service training mainly for secondary teachers (Li & Kwo, 2004).

Among the most significant developments in Hong Kong, teacher education was the establishment of the Hong Kong Institute of Education (HKIED) in 1994, which was renamed the Education University of Hong Kong (EdUHK) in 2016. The HKIED was formed by joining the five existing colleges of education at that time (Li &

Kwo, 2004). In 2000–2001, the HKIED joined the other seven tertiary institutions in Hong Kong under the financial, course validation and quality assurance procedures of the University Grants Council. In addition, the HKIED began to offer full-time four-year undergraduate programmes and part-time postgraduate diploma/certificate programmes for those of its former alumni who had, over the years, obtained an undergraduate degree through distance-learning and/or overseas degree programmes.

The entry requirement to a teacher education programme was low. Pre-service primary teachers either received two-year training after completing Form 7 (Year 13), or three-year training after completing Form 5 (Year 11), before they were enrolled in the HKIED for a Certificate of Primary Education. The degree courses were offered at HKU and CUHK, mainly for the purpose of building up specialist secondary teachers (Li & Kwo, 2004).

As teachers could become qualified through either sub-degree courses or degree courses, teachers were assigned to teach students of different levels according to the qualifications they held. Certificate holders who trained in the former colleges of education/HKIED taught mainly at primary and junior secondary level, whereas degree holders taught mainly at senior secondary level (Law, 2003). To cater for the expansion of education which occurred after waves of immigration from China before the late 1970s, the Hong Kong Government had to employ untrained teachers, which affected the quality of the teaching profession for decades. Teachers often had to teach more than one subject and had to teach subjects for which they were not trained (Law, 2003).

As stated above, in 2000–2001, the HKIED, the major teacher educator provider, joined the other seven tertiary institutions in Hong Kong for funding and quality control purposes. In addition, the HKIED increasingly began offering full-time four-year undergraduate programmes and part-time postgraduate diploma/certificate programmes for its former alumni who had, over the years, obtained an undergraduate degree through distance learning and/or overseas degree programmes. CUHK (The Chinese University of Hong Kong), HKU (The University of Hong Kong), and HKBU (Hong Kong Baptist University) also provided undergraduate and postgraduate programmes for pre-service and in-service teachers. The OUHK (Open University of Hong Kong, newly established in 1989), also offered degree courses and in-service and pre-service PGCE courses for primary and secondary school teachers (Lee, 2013).

As a response to the announcement made by the Chief Executive in 1997 that all future new teachers should be graduates and professionally trained, from 2002, all sub-degree courses, except in the area of early childhood education, were closed down (Morris, 2004).

Along with the increasing awareness worldwide that teachers should be both subject trained and professionally trained; a number of measures were taken in Hong Kong—both to enhance teacher professionalism and to gauge teacher professionalism. In 2000, Hong Kong's Education Commission launched a reform proposal entitled Learning for Life, Learning through Life (Education Commission, 2000). Following this report, a number of in-service training courses were provided in tertiary institutions. The language benchmark assessment (the LPATE) was also intro-

duced within this context, to make sure that all English and Putonghua language teachers met minimum language requirements (Li & Kwo, 2004).

Following these moves for greater professionalism, there was a considerable improvement in teachers' qualifications. In the year 2010–2011, of the 21,000 primary school teachers (including about 1600 non-degree holders), approximately 95% were trained; likewise of the 29,000 secondary school teachers (including 900 non-degree holders), about 94% were trained (Lee, 2013). As the discussion in Section II demonstrates the wide variety of language ability of teachers of different levels is invariably linked to the amount of academic and professional training teachers of English in Hong Kong have received.

It was also increasingly recognized by government that the days of non-graduate teachers were over if Hong Kong was to move on as a sophisticated, high-tech, service center with commensurate higher levels of education and language ability in its workforce.

Medium of Instruction in Schools

In Hong Kong, approximately 95% people are ethnic Chinese, most of whom have migrated from China's Guangdong Province. The remaining 5% come from places such as South Asia, East Asia, Europe, North America, or Australia (Census and Statistics Department, 2011). Despite the fact that the majority of people in Hong Kong spoke, read, and wrote Chinese, English was the sole official language until 1974. Chinese (i.e., Cantonese for the spoken language and Modern Standard Chinese for the written language) was recognized as an official language only after considerable pressure (Tsui, 2004).

The medium of instruction, i.e., using English as a Medium of Instruction (EMI), or using Chinese as a Medium of Instruction (CMI), has been debated since colonial times. Tsui (2004) points out that language policy is not solely an educational issue. The language policy must be understood in its social and political context. A review of the changes in the medium of instruction (MOI) in Hong Kong shows that the selection of medium of instruction is closely associated with the social, political and educational context of Hong Kong (see Jeon, 2016; Poon, 2013; Tsui, 2004).

Although English was the major medium of instruction in the colonial period, using Chinese as a medium of instruction was advocated as early as the 1960s, asserting that learning through a foreign language would impact negatively on the quality of learning, and that having a good foundation in the mother tongue was necessary for acquiring a second language (Tsui, 2004). In 1963, a government study about the educational needs of Hong Kong students showed that EMI education placed a heavy burden on students. The colonial government was nonetheless reluctant to support CMI. Indeed, a member of the Education Commission blatantly stated that anglicizing the Chinese would make them intermediaries between the colonial government and the local people (Pennycook, 1998; Tsui, 2004). On the other hand,

EMI education was seen as a means of satisfying parents in that it enabled students to better communicate with the international community.

In 1973–74, the Government once again proposed using Chinese as a medium of instruction but this proposal was not accepted: parental concerns and Hong Kong's economic development were put forward as the major issue blocking such a move. One change, however, was that the government left the choice of MOI to individual schools. For the first time, in 1974, the Hong Kong School Certificate Examinations could be taken either in Chinese or in English (Tsui, 2004). The change in the MOI was also associated with the fact that, in 1974, Chinese was recognized by the government as an official language.

Until 1994, the MOI was decided by individual schools, which Poon (2013) describes as the 'laissez-faire policy period'. Although schools claimed to be EMI schools, the mixed use of English and Chinese was prevalent in classes in these schools (Johnson, 1983; Lo & Lo, 2014). As only around 30% of students were able to learn through English effectively (Poon, 2013), most EMI schools used mixed-code teaching (a mixture of English and Cantonese) because of students' limited English language proficiency (Poon, 2013).

In 1994, the Hong Kong Government adopted a more rigorous language streaming policy. Schools were streamed into EMI schools, CMI schools, and two-medium schools on the basis of their students' language ability. Such a policy was not well received by parents and students, as the policy deprived schools of a free choice on the selection of the MOI (Jeon, 2016; Poon, 2013).

Upon the return of Hong Kong to China in 1997, the HKSAR Government introduced a compulsory CMI policy, whereby only schools with 85% of students achieving a satisfactory level of English over the previous three years would be permitted to use English as the MOI. Despite there being opposition from parents, such a strategy was supported by pedagogical evidence that learning through the mother tongue was more beneficial to students (Poon, 2013).

Since, for historical, ideological, and economic reasons, English enjoyed a high status in Hong Kong, many objections to the compulsory English policy were made by the public. Parents whose children now had to attend 'CMI' schools considered that such a policy deprived their children of access to higher education and good jobs (Poon, 2013). CMI school principals made the point that the policies made many CMI schools appear second class by limiting the number of high-quality students that CMI schools could enroll (Tsui, 2004). The CMI policy was also accused of restricting social mobility by blocking people's pathways to the elite (Poon, 2013). Despite the objections, research indicated that students benefited from learning in their mother tongue (see Marsh, Hau, & Kong, 2000; Ng, Tsui, & Marton, 2001). Nonetheless, despite such positive evidence, Hong Kong parents still showed an unwillingness to send their children to CMI schools.

As a response to some of the objections, the then Education Department issued the *Medium of instruction: Guidance for secondary schools* in September 1997—to permit schools to teach through the medium of English, provided that they demonstrated sufficient capacity to do so (<http://www.edb.gov.hk/en/edu-system/primary-secondary/applicable-to-secondary/moi/guidance-index.html>—accessed June 2016).

Schools that were permitted to use EMI would be subject to scrutiny every six years to ensure the quality of education; schools were also allowed to change their medium of instruction on the basis of student ability, teacher capacity, and the availability of support measures (Poon, 2013).

In 2010, the government amended the strict EMI policy, introducing ‘fine-tuning’ to the mix. Under the new framework, schools were permitted greater flexibility in deciding their medium of instruction. The fine-tuning policy allows a spectrum of MOI arrangements across schools, ranging from total CMI at one end, to CMI or EMI in different subjects in the middle, and total EMI at the other end. Under this policy, schools are allowed to offer EMI classes, partial EMI classes, or CMI classes based on students’ ability to learn through English, teachers’ capacities to teach through English, and school support (Jeon, 2016; Poon, 2013). Research into the fine-tuning policy has, however, reinforced many of the educational issues continually plaguing EMI. These issues concern whether students have sufficient language proficiency to study through a second language, whether teachers have the capacity to teach through English and whether sufficient resources and support are provided (Chan, 2014).

School Type

Originally, the majority of schools in Hong Kong were founded by religious bodies and merchant or clan groups. As education provision expanded, the government itself created schools that were directly funded from the public purse. Later, schools were founded by individuals or private bodies and funded from fees or funds provided by individuals. By the year 2000, government and religious/merchant schools were either directly resourced or ‘subvented’ by the HKSAR Government. It should be noted that the influence of school governing bodies, particularly religious ones, is very strong in Hong Kong. Indeed, it can be said that these bodies effectively set the curriculum in schools not the EDB.

Education Bodies and the Line of Command in Hong Kong

The policy bureau of the HKSAR Government is the Education Bureau (EDB).

The Hong Kong Examinations and Assessment Authority (HKEAA) is an autonomous body, established in 1978 to conduct all public examinations in Hong Kong. The Hong Kong Education Commission was an independent advisory body, established in 1982, in order to provide Government with policy advice (Coniam & Falvey, 2013).

Summary

This chapter has described the education and examination systems of Hong Kong. Chapter 5 describes the methodological approaches to the study and the analytical measurement tools used in the study.

References

- Adamson, B., & Li, S. P. T. (2004). Primary and secondary schooling. In M. Bray & R. Koo (Eds.), *Education and society in Hong Kong and Macao: Comparative perspectives on continuity and change* (pp. 35–60). Hong Kong: Comparative Education Research Centre, The University of Hong Kong.
- Benson, P. (2010). Teacher education and teacher autonomy: Creating spaces for experimentation in secondary school English language teaching. *Language Teaching Research*, 14(3), 259–275.
- Census and Statistics Department. (2011). *Population census*. Available at <https://www.censtatd.gov.hk/hkstat/sub/so170.jsp>.
- Chan, J. Y. H. (2014). Fine-tuning language policy in Hong Kong education: Stakeholders' perceptions, practices and challenges. *Language and Education*, 28(5), 459–476.
- Coniam, D., & Falvey, P. (2002). Does student language ability affect the assessment of teacher language ability? *Journal of Personnel Evaluation in Education*, 16(4), 269–285.
- Coniam, D., & Falvey, P. (2013). Ten years on: The Hong Kong language proficiency assessment for teachers of English (LPATE). *Language Testing*, 30(1), 147–155.
- Coniam, D., & Falvey, P. (2016). *Validating technological innovation: The introduction and implementation of onscreen marking in Hong Kong*. Singapore: Springer.
- Hong Kong Education Commission. (2000). *Learning for life, learning through life: Reform proposals for the education system in Hong Kong*. Hong Kong: Government Printer.
- Jeon, M. (2016). English language education policy and the Native-Speaking English Teacher (NET) scheme in Hong Kong. In R. Kirkpatrick (Ed.), *English language education policy in Asian* (pp. 91–111). Switzerland: Springer International Publishing.
- Johnson, R.K. (1983). Bilingual switching strategies: A study of the modes of teacher talk in bilingual secondary school classrooms in Hong Kong. *Language Learning and Communication*, 2, 267–285.
- Law, W. W. (2003). Globalization as both threat and opportunity for the Hong Kong teaching profession. *Journal of Educational Change*, 4, 149–179.
- Lee, J. C. K. (2013). Teacher education in Hong Kong: Status, contemporary issues and prospects. In X. Zhu & K. Zeichner (Eds.), *Preparing teachers for the 21st century* (pp. 171–187). Heidelberg: Springer.
- Li, T. S. P., & Kwo, O. (2004). Teacher education. In M. Bray & R. Koo (Eds.), *Education and society in Hong Kong and Macao: Comparative perspectives on continuity and change*. Hong Kong: Comparative Education Research Centre.
- Lo, Y. Y., & Lo, E. C. S. (2014). A meta-analysis of the effectiveness of English-medium education in Hong Kong. *Review of Educational Research*, 84(1), 47–73.
- Marsh, H., Hau, K. T., & Kong, C. K. (2000). Late immersion and language of instruction in Hong Kong high schools: Achievement growth in language and non-language subjects. *Harvard Educational Review*, 70(3), 302–347.
- Morris, Paul. (2004). Teaching in Hong Kong: Professionalization, accountability and the state. *Research Papers in Education*, 19, 105–121.

- Ng, D., Tsui, A. B. M., & Marton, F. (2001). Two faces of the reed relay. In D. Watkins & J. Biggs (Eds.), *Teaching the Chinese Learner* (pp. 135–160). Hong Kong: CERC, The University of Hong Kong.
- Nunan, D. (2003). The impact of English as a global language on educational policies and practices in the Asia-Pacific region. *TESOL Quarterly*, 37, 589–613.
- Pennycook, A. (1998). *English and the discourses of colonialism*. London: Routledge.
- Poon, A. Y. K. (2013). Will the new fine-tuning medium-of-instruction policy alleviate the threats of dominance of English-medium instruction in Hong Kong? *Current Issues in Language Planning*, 14(1), 34–51.
- Sweeting, A. (1990). *Education in Hong Kong pre-1841 to 1941*. Hong Kong: Hong Kong University Press.
- Sweeting, A. (1993). *A phoenix transformed: The reconstruction of education in post-war Hong Kong*. Hong Kong: Oxford University Press.
- Tang, K. C., & Bray, M. (2000). Colonial models and the evolution of education systems: Centralization and decentralization in Hong Kong and Macau. *Journal of Educational Administration*, 38(5), 468–485.
- Tsui, A. B. M. (2004). Medium of instruction in Hong Kong: One country, two systems, whose language? In J. W. Tollefson & A. B. M. Tsui (Eds.), *Medium of instruction policies: Which agenda? Whose agenda?* (pp. 97–106). Mahwah, N.J.: Lawrence Erlbaum Associates.

David Coniam is Chair Professor and Head of Department of the Department of Curriculum and Instruction in the Faculty of Education and Human Development at The Education University of Hong Kong, where he is a teacher educator, working with teachers in Hong Kong primary and secondary schools. His main publication and research interests are in language assessment, language teaching methodology, and computer-assisted language learning.

Peter Falvey is a teacher educator, formerly a Head of Department in the Faculty of Education, The University of Hong Kong. His main publication and research interests are in language assessment, first and second language writing methodology, and text linguistics.

Chapter 5

The Initial 1996 Consultancy Study



David Coniam and Peter Falvey

Abstract In December 1995, the Education Commission published Report Number 6 (ECR6), passing two issues to the Advisory Committee on Teacher Education and Qualifications (ACTEQ) for the latter's consideration and action. These were: (1) That minimum language proficiency standards should be met by all teachers in their chosen medium of instruction. (2) That levels of language and professional competence ('benchmark' qualifications) should be established for all language teachers. This chapter describes the initial 1996 benchmark consultancy study and what emerged from it.

Call for Tenders for an Investigative Consultancy Study

In early 1996, Education and Manpower Bureau (EMB) placed an advertisement in the Hong Kong press for tenders to investigate the establishing of benchmarks for teachers of English language, Putonghua and Chinese. The time frame was four months—from April to July 1996. For English, EMB proposed that benchmarks be investigated for the following purposes:

- To establish benchmarks for primary teachers/secondary teachers/tertiary educators
- To establish benchmarks for language teaching purposes/for promotional purposes
- To establish benchmarks for teachers of subjects other than English language (i.e. teachers of such content subjects as physics, history, mathematics) who use English as the medium of instruction.

The two editors of this book and their team were appointed to carry out the consultancy. The original consultancy team can be found in Appendix A “[Original Consultancy Team](#)”.

D. Coniam (✉) · P. Falvey
Department of Curriculum and Instruction, Faculty of Education & Human Development,
The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, Hong Kong
e-mail: coniam@eduhk.hk

© Springer Nature Singapore Pte Ltd. 2018
D. Coniam and P. Falvey (eds.), *High-Stakes Testing*,
https://doi.org/10.1007/978-981-10-6358-9_5

Composition and Objectives

The consultancy reflected a broad spectrum of expertise and experience in a number of relevant areas:

- Local and international *language teaching* experience and expertise at primary and secondary levels
- Local and international *language teacher education* experience and expertise as well as local and international *language educational assessment* experience and expertise.

The major objectives were two fold:

1. To investigate what and how prototype benchmarks might be established for lower secondary teachers of English
2. To investigate the kinds of test types and assessment instruments appropriate for determining prototype benchmark levels.

Once the consultancy study was underway, the team was asked to examine a number of further issues and to make recommendations. These additional issues included:

- Further trialling of benchmarks
- Implementation of the benchmarks and who should be benchmarked first
- Discussion of exemptions
- Discussion of enhancement courses for at-risk teachers—including course duration and mode of delivery
- Recruitment and training of assessors.

In addition to local teaching and teacher education experience, many of the members of the consultancy team also had experience of working with English language teachers on university courses. Their principal focus was to form reliable and agreed views of what would constitute a ‘minimum, agreed, acceptable standard’ for the ‘target language use’ situation (Bachman & Palmer, 1996, 2010) of an English language teacher of lower secondary level in the Hong Kong English language classroom.

Stages Followed in the Consultancy Study

Firstly, a review of the global literature was conducted on teacher assessment, teacher certification, performance tests and benchmarks.

The English language consultancy team then viewed and analysed videos which had been collected by the consultants in their years of working with and observing English language teachers. The purpose of examining videoed lessons was to define the underpinning constructs and skills that an English language teacher would be required to have/perform to a minimum standard in the target language use situation of the English language classroom.

In addition, the expertise of UCLES was drawn on. Its now-defunct Cambridge English Examination for Language Teachers, Level 1 (CEELT1), on which two of the consultancy team had worked, was a useful vehicle for considering the type of tasks that might be given to English language teachers as assessment tools.

The five-month process followed by the consultancy team can be summarised as follows:

- Constructs were identified and skills categorised.
- Prototype specifications were created for assessment instruments.
- A broad battery of tests was piloted.
- The battery of assessment instruments was amended, refined and extended.
- Scales (arising from the identification of constructs) and associated descriptors for the criterion-referenced benchmark instruments were developed.
- Investigations were conducted in order to set preliminary prototype benchmark levels (for lower secondary teachers of English).

In addition to the assessment instruments developed, survey data was collected at both local and international levels.

The purpose of the surveys was to:

- Investigate/review ‘benchmark’ or teacher certification patterns in other countries in the world
- Collect basic factual data about teachers of English language in Hong Kong
- Sample attitudes and beliefs of teachers of English language in Hong Kong.

The Hong Kong Survey

The match between the demographic data collected in the consultancy questionnaire and data obtained from a 1995 teachers’ survey conducted by the Hong Kong Education Department was a very close fit, indicating that the survey was very representative (see Coniam & Falvey, 1999 for details).

The consultancy team wished to investigate four areas: attitudes towards major aspects of benchmarking and standards—principally: language ability; subject-matter knowledge; pedagogic content knowledge; and the desirability of a professional teaching qualification.

The responses to the four questions set out in Table 5.1 reveal teachers’ beliefs and attitudes to the establishment of minimum standards.

In the four questions above, the overwhelming responses of the majority of English language teachers clearly indicated support for the establishment of agreed minimum standards in four key areas where benchmarks were being formulated. The responses indicated that there was widespread agreement for the establishment of minimum—standard language assessment.

As the sample size was large ($N = 9179$), even very small effects would show up as significant. Consequently, results in the table below are reported in terms of

Table 5.1 Responses to four key questions on teachers’ opinions on ‘minimum standards’

	Responses	Percent (%)	
Language ability: For teachers of English, I believe that there should be agreed minimum standards of <i>language ability</i> for English language teaching purposes			
Strongly disagree	60	0.6	↑
Disagree	97	1.0	1.6%
No opinion	1208	12.9	
Agree	5760	61.7	83.8%
Strongly agree	2063	22.1	↓
	9188	98.3	
Knowledge of the rules and systems of English: For teachers of English, I believe that there should be agreed minimum standards with regard to <i>knowledge of the rules and systems of English</i>			
Strongly disagree	33	0.4	↑
Disagree	131	1.4	1.8%
No opinion	1399	15.0	
Agree	6004	64.3	81.3%
Strongly agree	1585	17.0	↓
	9152	98.1	
General English language classroom teaching ability: I believe that, for teachers of English, there should be agreed minimum standards of <i>general English language classroom teaching ability</i>			
Strongly disagree	28	0.3	↑
Disagree	125	1.3	1.6%
No opinion	1254	13.4	
Agree	6410	68.6	83.0%
Strongly agree	1342	14.4	↓
	9159	98.0	
English language teaching qualifications: I believe that teachers of English should possess a recognised <i>teaching qualification in English</i>			
Strongly disagree	31	0.3	↑
Disagree	251	2.7	3.0%
No opinion	1715	18.4	
Agree	5878	62.9	76.9%
Strongly agree	1310	14.0	↓
	9185	98.3	

effect sizes rather than tests of significance. In line with Cohen (1988, pp. 477–478), differences are expressed in confidence intervals in terms of small (0.2), medium (0.5) and large (0.8) effects.

Table 5.2 presents the highest and lowest responses for the four ‘benchmark’ questions, together with the pooled standard deviations of all the groups being compared. A definite trend emerged in the responses to the four questions below that the responses from degree-holding secondary school teachers with a relevant professional qualification were different from the responses of non-degree-holding primary school teachers without a relevant qualification.

The contrast between the two groups was perhaps not surprising. Teachers most likely to agree with proposals for improvements to qualifications in the field tended to be those who chose to be in the field because they had a high degree of expertise and relevance in the subject they taught as well as a relevant qualification.

The ability band of the school in which teachers were teaching (see the description of school banding in the introduction to this section) made a difference to how they responded to the four major questions. There was a consistent tendency for those teaching in higher band schools, and holding a directly relevant degree, to prefer the establishment of language benchmarks as opposed to those teaching in lower band schools who did not hold a directly relevant degree.

In the context of the school’s medium of instruction, the greatest differences in the responses were between teachers possessing a relevant degree and teachers who had neither a degree nor a professional qualification in a related subject.

The largest differences between groups emerged in the context of years of experience. The groups most different were teachers with a relevant degree as opposed to those with a non-relevant degree. The Professional Teachers’ Union (PTU) reacted slowly to language benchmarks in 1997 and only carried out a small-scale survey. However, as reported in Chap. 7—where the process of establishing language benchmarks for primary teachers of English is described—the PTU eventually became much more vocal in its opposition to the mandatory imposition of language benchmarks for its serving teacher certificated (non-degree) members because they, more than any other group sampled, feared the outcomes of being assessed and judged.

Questionnaire on Language ‘Benchmarks’ for Teachers of English in Other Countries

On the issue of teaching qualifications, many countries now require their teachers to be qualified, although more so in secondary schools than in primary schools. Questionnaires were sent to 40 countries through the auspices of the British Council, where there is usually an English language teaching specialist who can comment with some authority on the English language teaching situation in that particular country. Twenty-one of the questionnaires sent out were returned (at 52.5%, quite a reasonable return rate), with the returns representing a cross section of the various parts of the

Table 5.2 Teachers' attitudes towards different types of 'benchmark'

	Degree: relevant secondary			Non-degree: non-relevant primary			Effect size
	Mean	SD	Cases	Mean	SD	Cases	
Attitudes towards a language ability benchmark	4.27	0.74	1494	3.94	0.61	3194	0.49
	WG	0.67	9179	WG	0.67	9179	(Medium)
Attitudes towards a subject-matter knowledge benchmark	4.18	0.70	1491	3.86	0.61	3186	0.49
	WG	0.67	9179	WG	0.67	9179	(Medium)
Attitudes towards a classroom teaching ability benchmark	4.14	0.68	1490	3.86	0.56	3188	0.45
	WG	0.67	9179	WG	0.67	9179	(Medium)
Attitudes towards the desirability of possessing an english language teaching qualification	4.11	0.71	1492	3.74	0.65	3199	0.54
	WG	0.67	9179	WG	0.67	9179	(Medium)

world—Latin America, Africa, Europe, Asia and the Middle East. While not exactly constituting hard data, the 21 responses nonetheless were broadly indicative of the general policy decisions of all the countries to which the questionnaires were sent.

Language benchmarks were reported to be in place in a number of countries. Some countries required not only secondary school teachers to have a minimum level of ability in English but primary school teachers as well. On the issue of teaching qualifications, it can be seen that the majority of countries required their teachers to be qualified, although more so in secondary schools than in primary schools.

On the question of whether English language teachers needed to have a subject-matter knowledge in English, i.e. that they must be graduates of English language/linguistics/applied linguistics/communications, 14 (70%) of the countries who responded indicated that their secondary school teachers of English needed to have a professional knowledge of English.

Test Battery

Teacher Sample for the Hong Kong 1996 Feasibility Study

Four groups of teachers—a total of 53 teachers—participated in the original consultancy study. One group consisted of pre-service teachers. The other three consisted of all the teachers of English language in three schools (see Table 5.3).

All teachers were videoed teaching one lower secondary class. Additionally, Groups 1, 3 and 4 were videoed as they took the newly developed oral tests.

Development of the Test Battery

In the construction of any assessment mechanism, it is useful if existing material is available for reference and, if possible, adaptation. Three members of the consultancy team had worked, at various times, on the construction of instruments for language teacher assessment at UCLES (Cambridge Assessment) and were familiar with various ability constructs and test types.

Table 5.3 Groups of teachers participating in the consultancy study

Group	Set of teachers
1	15 pre-service teachers
2	9 in-service teachers
3	16 in-service teachers
4	13 in-service teachers

While sections of the UCLES Cambridge Examination in English for Language Teachers, Level 1 (CEELT1) formed the backbone for the initial battery of tests, the battery of sub-tests (particularly the oral assessment and classroom language assessment) evolved over the course of the project as the consultants held meetings with the larger consultancy group. Specifically, two new test types for the oral examination instrument were trialled with the final school which participated in the project. The trial was successful, and the scales which evolved were successfully operationalised with the ACTEQ Language Benchmarking Task Force in June 1996. As will be described below, the battery underwent a number of changes through:

- Formative feedback while being administered to different groups of teachers
- Conceptual advice and assessment instrument moderation and feedback from the larger group of consultants.

This advice and feedback were particularly valuable in terms of:

- The constructs to be assessed in the different benchmarks
- Changes suggested for the oral test; in particular, with regard to the validity of the test
- Fundamental differences in assessment techniques and the results that are produced between a formal oral test involving replication and simulation tasks and language assessment in the live classroom. See the discussion in the literature review section (1.2) (c.f., Carlson, 1986).

An important issue involves the nature of classroom spoken discourse. Bachman (personal communication, 1996) suggested that an analysis of teachers' classroom discourse may reveal language traits that lend themselves—from an assessment perspective—to task types which may well be profitably developed into live classroom language test types. This was operationalised in the 1996 report (Coniam & Falvey, 1996) under *Task-Specific Specifications* when the final specification (the classroom oral language component) requires, under the 'input format', that teachers 'would be required to demonstrate language competence in presenting to and interacting with students'.

The data for the development of language benchmarks and the selection of appropriate task types for benchmark assessment, as mentioned, consisted of a battery of tests given to teachers and videos of lessons taught by teachers. Changes and amendments were made on the basis of the trialling, moderation, and the reconceptualisation and development of existing and new assessment instruments to assess the language ability of teachers and for creating benchmarks.

The different phases that the test types underwent as they were developed will now be described.

Phase 1—Initial Test Battery

The initial battery of tests consisted of a paper-and-pencil component comprising five sub-tests, an oral component which varied from two to four sub-tests and the observation of a live lesson. These tests were administered to Groups 1, 2 and 4 in the sample—a total of 37 test takers. The constitution of the test types was as follows:

1. Multiple-Choice Items and Reading Comprehension

This sub-test consisted of 54 short items and 3 reading comprehension passages. A total of 29 of the short items had been calibrated against a representative sample of the Hong Kong secondary school population (Coniam, 1995) and had also formed part of Hong Kong University's (HKU) and the Chinese University of Hong Kong's (CUHK) entry tests to their postgraduate certificate/diploma of education programmes. The remaining 25 items were from UCLES and were an anchor test calibrated against different groups on a worldwide basis.

2. Listening Comprehension

This sub-test was drawn directly from the UCLES CEELT1. It consisted of three passages, each based on a language teaching situation. Response types included short written responses.

3. Reading Comprehension

This sub-test was adapted from CEELT1. It consisted of one passage, based on a language teaching situation. Response types included short written responses.

4. Writing

This sub-test was taken from CEELT1. It consisted of one passage, based on a language teaching situation. Teachers had to respond to a written prompt.

5. Language Awareness

This sub-test, developed by Andrews (1999), consisted of 15 items in which teachers had to identify errors, explain the problem and correct them.

This sub-test was incorporated as a means of providing the consultants with comparative information about subject-matter knowledge. It was not intended to form part of the final battery of test types to be used in the assessment of language ability. It was envisioned that the results might be useful for reference if a subject-matter knowledge benchmark phase was ever to be initiated. However, as will be shown later, the rationale for this test type was subsequently amended and Writing Test Tasks 2a and 2b, which required teachers to identify and correct and then explain student errors, were introduced much later in the development of language benchmarks as tests of teacher-specific language skills (see p. 100 "The Writing Test".)

Table 5.4 Scales and constructs identified for classroom language assessment

Constructs	Assessment scales
<ul style="list-style-type: none"> • The ability of teachers to use English in the formal section of the lesson when presentation and practice takes place 	The language of presentation/practice
<ul style="list-style-type: none"> • The ability of teachers to use English to interact with students during the language lesson and elicit meaningful responses 	Interaction with students
<ul style="list-style-type: none"> • The ability of teachers to use English accurately in the language classroom 	Grammatical accuracy
<ul style="list-style-type: none"> • The ability of teachers to pronounce English accurately in the language classroom 	Pronunciation, stress and intonation

6. Oral Tests

This component originally consisted of two sub-tests which were taken directly from CEELT1.

(i) *Reading aloud (on a one-to-one basis with the examiner)*

There were two sub-tests in this part of the oral test. The first sub-test involved teachers reading an extract from a narrative text. The second sub-test involved teachers in giving instructions to a class.

(ii) *Group interaction (in groups of 3/4 teachers)*

The group discussion task was based upon viewing a video clip of a teacher teaching a class.

7. Videoing of a Live English Language Lesson—The ‘Classroom Language Assessment’ Component

The consultancy team considered that this was the most important part of the test battery, since it represented a performance test of the actual target language use situation. Its development and validation is described in depth in Coniam and Falvey (1999). Initial discussion centred on an analysis of numerous videos of English language teachers teaching English language classes. From this analysis, four constructs and their scales emerged as the test bed for the Classroom Language Assessment component (see Table 5.4).

The original intention had been to produce scales and descriptors which were dichotomous. Teachers were simply ‘benchmarked’/‘not benchmarked’. After much discussion, however, and to allow for future potential uses (e.g. levels above the benchmark being used for purposes of promotion and substantiation and levels below the benchmark being used for constructive feedback), a five-point scale was produced, with the mid-point, i.e. a ‘3’, set as the tentative benchmark level.

5	Complete ability	
4	Extensive ability	
3	Moderate ability	← benchmark level
2	Limited ability	
1	No ability	

Phase 2—Amended Test Battery

Two features differentiated the amended battery of tests from the initial battery. Firstly, the *language awareness* sub-test was dropped. Secondly, substantial revision was conducted on the oral tests. The reconstitution of the test types was as follows:

1. Multiple-Choice Items and Reading Comprehension

As for the original (i.e. Phase 1) test battery.

2. Listening Comprehension

As for the original test battery.

3. Reading Comprehension

As for the original test battery.

4. Writing

As for the original test battery.

5. Language Awareness

Dropped.

6. Oral Tests

There were four sub-tests in the amended battery of oral tests:

- (i) ***Reading aloud (on a one-to-one basis with the examiner)***
The second of the original CEELT1 reading aloud passages was retained (giving instructions to a class).
- (ii) ***Explaining and interpreting a lesson (on a one-to-one basis with the examiner)***
Having studied a transcript of an intended lesson together with an accompanying flow chart, this activity required teachers to explain to a third party how the lesson would be taught using the flow chart only.

Note: The lesson plan concept was originally trialled in the format of a flow chart. While the flow chart was found to be effective in stimulating teacher talk about teaching, on the advice of the other consultants, it was agreed, for purposes of authenticity, to modify the input to resemble a more conventional lesson plan rather than a flow chart (see Appendix B “[Talking about Teaching Subtest](#)”).

(iii) ***Oral interaction based on language problems (in groups of 3 or 4)***

This activity required teachers to identify errors in sentences, correct them and discuss, in the group, methods and techniques to be used in class to remedy the problems.

The stimulus was found to be effective in provoking teacher discussion and interaction. However, as a result of feedback from the other consultants and from members of the ACTEQ Language Benchmarking Task Force, it was agreed that a more suitable stimulus, in terms of washback effect on teachers, would be an authentic sample of connected prose containing typical student errors. The original stimulus is provided in Appendix C “[Student Essay for Discussion of Errors](#)”.

(iv) ***Group interaction (in groups of 3/4 teachers)***

A discussion (retained from the original battery) based upon viewing a video clip of a teacher teaching a language class.

As a result of the successful trialling of the two new sub-tests, it was decided to drop the video-based oral interaction component (i.e. [iv] above) from the tasks recommended to the ACTEQ Language Benchmarking Task Force.

7. **Videoing of a Live English Language Lesson**

This element was retained unchanged from the initial test battery.

Consultancy Study: Recommendations on *Language Ability* Benchmarks

There were a number of objectives for the consultancy study. The first of these was how language benchmarks for lower secondary English language teachers might be formulated, and the type of test assessment instruments that might be appropriate to use with English language teachers. In addition, the consultancy team was asked to consider the issue of training programmes for teachers—length, course provider constitution, selection and evaluation. The consultancy team was also aware that the consultancy study had focused solely on benchmarks for lower secondary English language teachers. If the initiative was to have credibility, it would need to encompass the range of levels at both primary and secondary level. While the issue of language was an important one for an English language teacher, it was also important not to lose sight of subject-matter knowledge and pedagogic content knowledge. (It will be recalled that English language teachers possessing a relevant degree and a

professional teaching qualification constituted less than 20% of the secondary school English language teaching cohort at that time.)

The consultancy team was nonetheless mindful that the consultancy study had to be viewed in the light of an initial exploratory study. The study had been a small-scale investigative feasibility study where the research question was:

What approaches might be taken to investigate the feasibility of establishing minimum language standards for English language teachers rather than actually setting minimum language standards?

On this basis, while it was important to make recommendations to ACTEQ, it was also important that Government not take it for granted that *setting* minimum language standards had been either attempted or accomplished.

In the section below, discussion focuses on the different recommendations that the consultants made to ACTEQ. These involved:

- The constitution of appropriate assessment instruments
- Enhancement programmes—potential length, methods of evaluation
- Other areas to be benchmarked
- The way forward.

Recommendations on the Test Battery

On the basis of teachers' responses to the questionnaire, it was clear that the vast majority of teachers believed that minimum standards for language ability should be a prerequisite for all teachers of English. On the basis of the in-depth case studies and extensive consultation, it was recommended that in establishing language ability benchmarks, consideration should be given to benchmarking English language teachers in the following areas:

1. Reading
2. Writing
3. Listening
4. Speaking
5. Classroom Language Assessment.

It was recommended that the first three areas above be assessed in a formal examination setting. The fourth area (Speaking) would be assessed in an interview situation and the fifth area (Classroom Language Assessment) should be assessed *directly*, in a live classroom setting.

At this stage, however, the benchmark initiative was still very much prototypical. Definite recommendations for the manner in which those tests might be implemented as final versions were *not* included. Rather, for each test type, a number of general and specific recommendations were made for ACTEQ and the teacher education and assessment community to consider.

Specifications for Benchmark Tasks for the Assessment of Language

In criterion-referenced assessment, it is important that the reliability of the assessment tasks is ensured, particularly in a high-stakes form of assessment. It is important that tasks which are run in parallel (e.g. in multiple assessment administrations) are as comparable as possible. The consultancy team thus strongly recommended that rigorous task specifications be used by those appointed to set tasks and applied by moderation committees once the task specifications had been finally agreed and formally ratified.

Although the creation of task specifications was not part of the remit for the 1996 consultancy, draft prototype task specifications were provided as a guide to the benchmark administration body.

It was noted that the prototype task specifications for each of the five areas to be benchmarked were tentative and would need to be refined over the following months so that exemplar tasks could be provided for and trialled on teachers who would attend the pilot assessments in 1997–1998.

Reading and Listening

At this stage, it was necessary to comment on the differences between methods of assessing speaking, classroom language and writing compared to the assessment of reading and listening.

Speaking, classroom language and writing are observable instances of human interactive behaviour. They are sometimes described as ‘productive’ abilities. While this may be something of an imperfect definition, it is useful to distinguish between an activity which, in itself, is a product or performance (text or speech) against an activity which is not (reading and listening).

The usual methods for assessing reading and writing are:

- Indirect methods such as paper-and-pencil tests of comprehension
- Methods in which reading and listening are linked to a related activity such as speaking or writing. Such an activity, created to provide an opportunity to demonstrate understanding, is referred to as *integrated assessment*.

In the initial stages of the benchmarking project, it was suggested that both methods (paper-and-pencil and integrated assessment) be trialled in the pilot assessments for teachers of English. The paper-and-pencil tests provide a score. When that occurs, a cut score has to be decided upon (see Appendix “Methodological Approaches and Analytical Tools” in Chap. 8 for an explanation of cut scores). This is neither helpful nor informative because there are neither rating scales nor descriptors which teachers can use for self-assessment or which course providers can use for diagnosis and feedback. The integrated tests, however, lend themselves to the creation of scales and

descriptors which have all the advantages described above. The consultants considered recommending that research be conducted into the creation of rating scales and descriptors for reading and listening by the Hong Kong-based Standing Committee on Language and Research (SCOLAR). Such research would entail investigating whether rating scales and descriptors based on a hierarchical model of reading and listening skills (e.g. from the lowest, such as recall and recognition to higher order reading/listening skills involving synthesis and/or analysis, and/or evaluation) could be developed as a feasible assessment instrument.

In the meantime, the recommendations for reading and listening were that both paper-and-pencil and integrated tests be used on the pilot assessments and compared with the results obtained by test takers on the other sub-tests.

Task Specifications (General and Specific)

It was suggested that there should be both general specifications (applicable to many tasks) and specific specifications for selected tasks.

General Specifications

For benchmarking tasks for the Reading, Listening, Writing Tests and parts of the Speaking Test components, the texts used should be both authentic (see Bachman & Palmer, 1996) and related to English language teaching/learning.

Task-Specific Specifications

In the 1996 consultants' report (Coniam & Falvey, 1996), it was made clear that the specifications shown below were neither exclusive nor fixed. They demonstrate types of specification documentation that are required to help task creators produce comparable tasks and task moderators to moderate them. They are mainly based on exemplar tasks which were used or created for the collection of data during the consultancy. Originally provided for benchmark administrators, they are included here in order to provide readers with a guide to task production.

1. The Writing Test

The UCLES CEELT test was very useful in providing a basis for what was subsequently developed for the Writing Test.

Title	Writing
Purpose	To assess teachers' ability to write connected prose in a writing task which simulates writing tasks they may have to perform as part of their job
Constructs to be assessed	1. Organisation and coherence 2. Grammatical accuracy 3. Task completion
Duration of task	30 min
Length of task	300 words
Input format	Written stimulus, describing the expected response in terms of content, register and audience
Scoring procedure	Application of scales with associated descriptors
Test type	Writing connected prose
Benchmark for lower secondary teachers of English	A score of '3' on each of the three scales, viz.: 1. Organisation and coherence 2. Grammatical accuracy 3. Task completion

Appendix D “[Writing Test Scales and Descriptors](#)” presents the initial version of the Writing Test scales and descriptors.

2. The Reading Test

The CEELT Reading Test had face validity because reading comprehension is assessed within the general context of English language teaching and teachers are familiar with it as an assessment instrument. The consultancy team felt some concern, however, over a CEELT-type test which simply assessed reading comprehension in terms of a ‘read-the-passage-and-answer-the-questions’ traditional framework. The English language public examinations in Hong Kong have moved towards a more communicative framework where reading and listening are assessed, to an extent, within an integrative framework. For teachers to be assessed in a discrete mode would be a retrogressive move, it was felt. A number of recommendations were submitted to ACTEQ for consideration. These included specifications for assessing reading in a traditional framework as well as recommendations for further investigations into how reading might be assessed in a more integrative manner. Two sets of specifications are provided below.

Title	Reading
Purpose	To assess teachers' ability to read connected prose and respond to multiple-choice questions
Language focus	Understanding of professional text
Duration of task	30 min
Input format	Written text of approximately 500 words, taken from an authentic English language teaching source with written questions requiring a response which demonstrates an understanding of the text
Scoring procedure	Correct/incorrect
Test type	Multiple-choice
Benchmark for lower secondary teachers of English	A minimum score—to be decided once scores had been matched against the profile of a large enough sample of typical benchmarked teachers

Title	Displaying understanding of a text
Purpose	To assess teachers' ability to read connected prose and demonstrate understanding through the production of notes/connected text
Language focus	Understanding professional text
Duration of task	30 min
Input format	Written text of approximately 500 words, taken from an authentic English language teaching source
Scoring procedure	Application of scales with associated descriptors
Test type	Writing notes/connected text in response to questions testing understanding of an authentic written text about teaching
Benchmark for lower secondary teachers of English	A '3' on the five-point scale

3. The Listening Test

The Listening Test faced problems similar to those of the Reading Test. The CEELT Listening Test, which was used in the 1996 consultancy study, was based on a text set in the context of English language teaching. The consultancy committee, as for the Reading Test, felt that a traditional 'listen-and-answer-the-questions' framework was inappropriate, for reasons similar to those expressed above about the Reading Test. A number of recommendations were therefore submitted for ACTEQ to consider.

Title	Listening and reporting
Purpose	To assess teachers' ability to listen to and understand teachers talking about language to fellow teachers
Language focus	Understanding extended spoken discourse
Duration of task	20 min (listening task of approximately 6–8 min)
Input format	(i) Tape recording/video with main points to be relayed by test takers to peers (ii) Oral instructions from interlocutor
Scoring procedure	Application of scales with associated descriptors
Test type	Listening to a taped dialogue and demonstrating understanding by reporting and responding to prompts from an interlocutor
Benchmark for lower secondary teachers of English	A '3' on the five-point scale

Title	Listening and responding
Purpose	To assess teachers' ability to listen to and understand teachers talking about language to students
Language focus	Understanding extended spoken discourse
Duration of task	20 min (listening task of approximately 6–8 min)
Input format	(i) Tape recording/video of teachers talking to teachers (ii) Oral instructions from interlocutor
Scoring procedure	Application of scales with associated descriptors
Test type	Listening to a taped dialogue and demonstrating understanding by reporting and responding to prompts from an interlocutor
Benchmark for lower secondary teachers of English	A '3' on the five-point scale

4. The Speaking Test

As outlined above, the Speaking Test underwent substantial change and development throughout the course of the consultancy study. The consultancy team felt that the major elements trialled for the Speaking Test were appropriate.

Title	Speaking—Reading aloud
Purpose	To assess the teachers' ability to pronounce effectively and communicate with an audience when reading aloud
Language focus	Pronunciation
Duration of task	2 min—preparation 3 min for task completion
Input format	Text of approximately 300 words to be read aloud
Scoring procedure	Application of two scales with associated descriptors
Test type	Oral interaction—simulation of classroom activity
Benchmark for lower secondary teachers of English	A score of '3' on each of the two scales, viz.: 1. Pronunciation, stress and intonation 2. Communication of text to audience

Title	Speaking—Talking about teaching
Purpose	To assess the teachers' ability to use English for the purpose of explaining and interpreting a lesson
Language focus	Pronunciation; grammatical accuracy; organisation of spoken discourse
Duration of task	5 min—preparation 6 min for task completion
Input format	(i) Transcript of spoken text describing the lesson to be discussed (ii) Lesson plan of the lesson to be discussed (iii) Written rubric for the task
Scoring procedure	Application of three scales with associated descriptors
Test type	Using written stimuli (transcript and lesson plan) to interact with an interlocutor
Benchmark for lower secondary teachers of English	A score of '3' on each of the three scales, viz.: 1. Pronunciation 2. Grammatical accuracy 3. Organisation and cohesion

Title	Speaking—Professional oral interaction
Purpose	To assess teachers' ability to use English for purposes of professional oral interaction
Language focus	Language of interaction with professional peers
Duration of task	5 min—preparation 15 min for task completion
Input format	(i) Written stimulus in the form of authentic student written text containing typical student errors (ii) Oral instructions from interlocutor
Scoring procedure	Application of two scales with associated descriptors
Test type	Sharing the same piece of authentic written text with two peers to demonstrate the ability to interact with peers in the process of explaining professional matters
Benchmark for lower secondary teachers of English	A score of '3' on each of the two scales, viz.: 1. Interacting with peers 2. Explaining language matters to peers

Appendix E “[Speaking Test Scales and Descriptors](#)” presents the initial version of the Speaking Test scales and descriptors.

5. Classroom Language Assessment

The test type which created most discussion in the consultancy study committee was the Classroom Language Assessment (CLA). In part, this was because it was a ‘strong’ performance test (see, e.g., McNamara, 1996). It would not only be seen to be directly relevant to English language teachers’ work in the English language classroom, it would also have considerable washback effect. A direct test such as the CLA which would be conducted by assessors, travelling around Hong Kong, visiting teachers in live classes would be very expensive for Government. In addition, for issues of CLA reliability as an assessment instrument, double-assessment would be required.

Title	Classroom language assessment component
Purpose	To assess teachers' ability to use English grammatically and with appropriate pronunciation for the purpose of interacting with students and formal presentation/exposition
Language focus	Language of presentation/exposition; language of interaction with students; grammatical accuracy; pronunciation
Duration of task	A whole lesson
Input format	Teachers would be required to demonstrate language competence in presenting to and interacting with students
Scoring procedure	Application of four scales with associated descriptors
Test type	A sample of authentic language used by the teacher in a classroom environment
Benchmark for lower secondary teachers of English	A score of '3' on each of the four scales, viz.: 1. The language of presentation/practice 2. Interaction with students 3. Grammatical accuracy 4. Pronunciation, stress and intonation

Appendix F “[CLA Scales and Descriptors \(from Consultancy Study\)](#)” presents the initial version of the Classroom Language Assessment scales and descriptors.

Benchmark Level Recommended

At this stage of the investigation, it was recommended that a benchmark level of ‘3’, described as ‘moderate ability’ on each scale, should be used to indicate minimum competence for the following components:

- The Speaking Test component
- The Writing Test component
- The Classroom Language Assessment component
- Integrated tasks for the Reading and Listening Tests which would allow for the creation of appropriate constructs, scales and descriptors.

Other Recommendations

Gaining Broader Acceptance of the Benchmarks and Benchmark Tests

It was emphasised that the consultancy study was a feasibility study and that trialling on a much larger scale would be needed before benchmarks could be finalised or ready for implementation as policy.

A major recommendation therefore was that a representative panel be constituted to examine the report of the consultancy study and to consider ways in which the benchmark initiative might be moved forward. It was recommended that the panel be constituted as a Subject Committee under the Hong Kong Examinations Authority.

Estimating the Amount of Training Required for the English Language Teaching Force

This was not possible without a larger scale study. A representative pilot study would hopefully provide a more reliable estimate of what proportion of the English language teaching force may require training.

Establishing Language Development Programmes

Among the issues discussed above in Chap. 3, issue 3.10 states that the provision of opportunities for test takers to attend upgrading courses (or enhancement programmes) is an essential requirement of any high-stakes form of assessment. The

consultancy study report recommended seeking indications of interest from potential providers of upgrading/development programmes such as the tertiary institutions and private bodies, after the publication of the report.

At this point, it is important to note that the benchmark initiative had not been conceived merely as a test for teachers to pass, be benchmarked and permitted to continue teaching. Nor was it conceived that failure would lead to disbarment from teaching English. One very important aspect of the Hong Kong benchmark initiative was the allocation of substantial resources for teachers at risk to enrol on upgrading courses. In November 2000, the HKSAR Government announced that HK\$240 million (approximately US\$30 million) had been set aside for enhancement programmes—an allocation of HK\$13,000 (US\$1700) per teacher of English. After consideration by the English Language Benchmark Subject Committee (ELBSC), it was recommended that once benchmarks had been finalised, courses of between 100 and 200 h duration be set up.

Criteria for Selection of Test Takers for the Pilot Programme

It was recommended that the following categories of teachers be invited to participate in the pilot programmes:

- New teachers of English who did not possess either English as a major subject in their first degree and/or who did not possess an appropriate teaching qualification
- In-service teachers of English who had less than 3 years' service and who did not possess either English as a major subject in their first degree and/or did not possess an appropriate teaching qualification
- In-service teachers of English with more than 3 years' experience who did not possess a teaching qualification/had not attended any refresher courses
- Those seeking confirmation of appointment as teachers of English
- Those wishing to be considered for promotion
- Volunteers (the 1996 questionnaire asked for volunteers willing to be assessed and videoed in the classroom).

Priority as to Who Should Be Benchmarked First

The decision who to benchmark first was a policy decision to be taken by the HKSAR Government. However, by mid-1996, it appeared logical to benchmark the following categories of teacher first:

- Those new to the profession, i.e. teachers in pre-service training
- Those seeking either confirmation of employment or promotion
- Those possessing neither formal subject knowledge nor a formal teaching qualification.

Implementing the Classroom Language Assessment Component

Since language ability in the live classroom situation may differ from language ability in a controlled situation (Coniam & Falvey, 1998), it was considered important to retain this element within the benchmarking initiative. To exclude CLA from the suite of benchmarking instruments would remove the opportunity to assess spoken language both in the classroom and at the higher level of interacting with peers, it was argued.

It was recommended that various options for operationalising this benchmark should be considered, as follows:

1. The use of video-recorded lessons, to be assessed by trained assessors
2. A 'cascade' scheme whereby a group of trained assessors train other groups who train further groups responsible for a limited number of schools
3. CLA assessment to be carried out by the Advisory Inspectorate of the Government's Education Department
4. Cross-assessment by 'benchmarking' panel chairs of English (panel chairs would not, in this option, assess teachers in their own schools).

It was recommended that the various methods of operationalising this benchmark should be the subject of consultation and discussion between ACTEQ, Education Department, school principals and teachers and other professionals in the field. The consultants believed that options (2) and (4) would provide the best opportunity for minimising problems, lowering costs and maximising the washback effects of implementing this benchmark.

Recruiting and Training Assessors

In order to ensure that reliable standards of assessment could be established and maintained—particularly for the speaking and classroom language benchmarks—it was recommended not only to use two assessors for CLA but also, in the long run, to build up a cadre of benchmark assessors who are regularly exposed to refresher standardisation sessions.

It was recommended that the body which administers the benchmarks would eventually be responsible for recruiting, training, standardising and updating assessors. In the short term, the consultants agreed to train the first batch of assessors.

Exemptions/Renewal of Benchmark

In relation to issue 3.10, decisions about who to exempt and for what reasons were policy decisions to be decided by the HKSAR Government. The options available were either one of blanket exemptions for certain categories of teacher or an examination of cases for exemption on a case-by-case basis. However, in 1996, it was not felt possible at that time, either to make recommendations on exemptions or recommend the length of time that a benchmark might remain current.

Professional Body for Teachers

It was generally agreed that the establishment of such a body was vital for the well-being of teachers' professionalism with a code of conduct, a code of ethics and the ability to monitor and upgrade the professional abilities and knowledge of its members.

Summary

This chapter has discussed the background to the initial consultancy and has shown how the consultancy investigated the feasibility of benchmarks, and amended and moderated initial assessment instruments. At the conclusion of the initiative feasibility study, the then-Education and Manpower Bureau proposed the establishment of an English Language Benchmark Subject Committee (ELBSC), which would be constituted and conducted in ways similar to other subject committees of the Hong Kong Examinations Authority (HKEA). The role that the ELBSC took in developing language benchmarks is now considered further in Chap. 7.

Appendix A: Original Consultancy Team

	Position	Institution
<i>Principal investigators</i>		
Dr. David Coniam	Professor	Department of Curriculum and Instruction, The Chinese University of Hong Kong
Dr. Peter Falvey	Senior Lecturer and Head of Department	Department of Curriculum Studies, The University of Hong Kong
<i>Consultant investigators</i>		
Dr. Stephen Andrews	Lecturer	Department of Curriculum Studies, The University of Hong Kong
Prof. Lyle Bachman	Chair Professor and Director Professor	English Language Teaching Unit, The Chinese University of Hong Kong Department of TESOL and Applied Linguistics, University of California at Los Angeles (UCLA)
Ms. Ann Cheung Yuet Yau	Lower-form Panel Chair	Immanuel Lutheran College, Tai Po
Dr. Jenny Chung Sing Ling	Lecturer	Hong Kong Institute of Education
Ms. Annie Ho Siu Wah	Senior Education Officer (Administration)	Vocational Training Council
Ms. Christina Lee Wong Wai	Subject Officer (English)	Hong Kong Examinations Authority
Dr. Michael Milanovic	Director of ELT Division, Head of Testing and Development Unit	University of Cambridge Local Examinations Syndicate, Cambridge, UK
Mr. Roderick Pryde	Director	English Language Teaching Institute, The British Council, Hong Kong
Dr. Sima Sengupta	Teaching Consultant	Department of Curriculum Studies, The University of Hong Kong

Appendix B: Talking About Teaching Sub-test

(Preliminary data collection task)

Talking about Teaching

Time for preparation: 5 min

Time for task completion: 6 min

Instructions

The following text is the words of a teacher talking about the lesson she is going to be teaching. An outline of her lesson plan (matching her explanations) of what will happen in the lesson accompanies the text. Read through the teacher's description so that you understand the outline in order to complete the task below

Task: Your task is to explain this lesson to a colleague (the Interlocutor) using the outline as the basis for your explanation. Note that while you need to understand what will happen in the lesson, and you need to understand the outline, *you do not need to memorise the teacher's explanation of exactly what she will be doing*. Try to explain what the teacher is going to do (in terms of the steps she is going to take, and her aims and materials) in your own words. You will not be allowed to retain the text while explaining the outline

Jane's Description of her Lesson

"OK, let me talk you through what I'm going to do in my lesson and how I think it's going to go. As you can see the lesson is about describing objects. The focus for the lesson will be about lost objects - reporting something you've lost. This will then lead up to a final piece of listening - although I haven't worked this out in detail yet. So my objectives, I would say, are a mix of the language necessary for describing objects and listening for specific information.

The first part of my lesson will be a lead-in. Here, I'm going to ask students if they have ever lost anything. Usually somebody's always lost something, and this helps to bring out useful early vocab like 'wallet' 'ID card' etc., and what you do if you have lost something - like reporting it at the school office, or having to go to the police station. Whatever useful words I get from them I'll write up on the board and have all the class repeat.

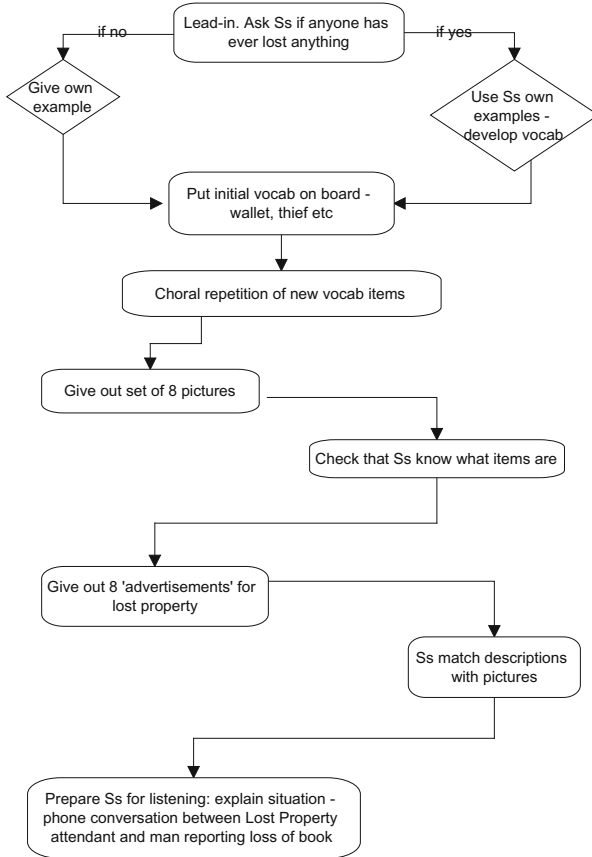
Then I've got a set of 8 pictures that I'll give out where I'll ask students to first check that they know what the name of the objects are - 'scarf', 'video camera' etc. Then I've got some short descriptions of lost objects - you know, the sort of thing that you might stick on the notice board in your block of flats to advertise that you've lost something. In pairs, students have to match the descriptions to the pictures.

Next I'll come to the listening. I think that I'll use a dialogue - a telephone conversation - between a Lost Property Office attendant and a man who's lost a library book, although this is as far as I've got and I haven't really thought it through yet."

*OBJECTIVE: Describing objects (personal effects)
so that you can recognise something you have lost*

AIM

MATERIALS



Warm up - get Ss' ideas

Consolidate what Ss may have offered

pictures of objects

Tune Ss in to pictures

Introduce descriptions of lost objects

short written descriptions

Warm Ss up for

tape

Appendix C: Student Essay for Discussion of Errors

Task: Read the composition and correct the errors. Identify errors in one paragraph and discuss possible ways of correcting them. Comment on positive and negative aspects of the student's writing.

Sample composition 1 (Topic: *My ideal school*)

My ideal school have been that item: included air condition for each study room, the canteen provide good, cheap and variety food, regular-meetings between student representatives and teachers and more emphasis on sports and other extra-curricula activities.

First of all, I need my school have a air condition for each study room. When the summer is coming, the whether will be getting hot. Especially the student after Physical Education, they would be getting very hot. At this moment the air condition can available. If without the air condition the student may be lost the interest in another lesson after physical education. Some time the cold air can make a man clear. The canteen always provided some bad taste food for the student. In addition, the food price also very expensive, but no another choice for him. So the student must to be ate these bad taste food. If the school can provided more variety food for the student make his own taste will be better.

My ideal school should have a regular meeting twice a week between student representatives and teachers. I discover that some teacher as well as teach the student. They lacked a communication with the students. I feel that when the teach more contact the student, he will be liked his lesson more. If have a regular meeting, they have the chance to communicate for student and the student can asked the teacher advice.

Finally, I need my school have more emphasis on sports and other extra-curricular activities, because many teenages like the sport. So we need more extra-curricular, supply to him.

Sample composition 2 [Topic: Next month you are going on a summer camp to Sai Kung with some of your friends. Choose two or three of the activities shown below that the camp offers, say why you want to do them and how you will prepare for them. (Pictures showing canoeing, rock-climbing, fishing, barbecuing and cycling)]

I am going on a summer camp to Sai Kung with my friends next month. Here are many activities for choosing by me. I will choose the cycling and the barbecuing.

Cycling is a good sport. I always rid it before I come to Hong Kong. However, I have never played it a long time. I like it very much so that I use the chance to play it. Cycling can train my body to get a healthy. Moreover, it can make me gaining adequate strength. As I am playing it, it can bring me a satisfactory and exciting feeling. Furthermore, I may see the beautiful sight along the road.

Barbecuing also is a interesting activity. The foods which passing are barbecued that it is a good food. In addition, barbecuing can increase friendship of my friends and I. On the other hand, it can teach me much knowledge which can't learn at the school. In view of these reasons, so I choose them to play.

Before I play them, I will prepare adequate for them. In the first place, I shall rent a bicycle from the bicycle shopping. Then, I will play it slowly at the garden or

park because it is necessary to train a good skill. Besides, I will find the experience about barbecuing from the books because it is my first time to barbecue. Finally, I will prepare adequate facilities about them since I must get safely and happy at the trip.

Sample composition 3 (Topic: as for composition 2)

I will go on a summer camp to Sai Kung with some of my friends next month. I think I will learn a lot of things during the camp, because I had chosen three activities to learn, including: canoeing, fishing and cycling.

I chose these because I don't know how to play them before. I always look another people to play, because I can't play them. Therefore I must learning how to play them well.

Before I go on this camp, I must prepare all the things what I will need. I had bought a cloth for swimming. My father made a stick for me to learn fishing. Finally, I had bought some T-shirt and jeans, I know these clothes will suitable for cycling.

Apart from theses things, I must bring so medical and some bandages. If we have any accident, they will save us. Beside these, I will bring some sun oil, too. Because I will learn canoeing and fishing at the sea. So, the sun must very bright and it will hurt my skin.

I had prepared all things which I will need. So, I hope I will have an excited and wonderful summer camp.

Appendix D: Writing Test Scales and Descriptors

	Task 1: Expository writing	Scale 2: Grammatical accuracy	Scale 3: Task completion	Task 3: Rewriting student composition	Scale 5: Organisation and presentation of facts/information
Well above the BM	The development of the ideas is smooth, logical and easily accessible to the reader. A superior piece of writing. Where necessary, propositions are justified and/or elaborated or illustrated with examples to enhance meaning. The writer has produced a highly coherent piece of text. Displays full audience and appropriate register	Grammatical structures are always accurate, with no occurrence whatsoever of non-idiomatic or other inappropriate expressions. There is access to a wide range of structures, which can be invoked at any time. Any 'mistakes' that occur can be categorised as lapses rather than systematic errors	All of the content demanded of the writer by the task is present. The task is fulfilled in an exemplary manner. A superior piece of writing. The writer displays an extremely high degree of sensitivity to the audience to produce a very competent piece of text	Incorporates a wide range of additional appropriate expressive vocabulary. Accurately corrects grammatical errors using a wide range of appropriate grammatical forms. Switches from one grammatical form to another accurately. Skilfully adds cohesive elements to text. Resynthesises information with extended vocabulary making meaning more explicit	Uses appropriate, contextualised and grammatically correct opening paragraph to 'orientate' or give background information to the reader. Uses headings where appropriate. Uses a style and tone (register) appropriate for purpose of writing. Includes and fully organises all facts/information from the original, possibly with some expansion; introduces new paragraph(s) where appropriate. Completes task, appropriate correct ending

(continued)

(continued)		Task 1: Expository writing			Task 3: Rewriting student composition	
	Scale 1: Organisation and coherence	Scale 2: Grammatical accuracy	Scale 3: Task completion	Scale 4: Vocabulary and grammar	Scale 5: Organisation and presentation of facts/information	
Above the BM	The logical flow of the text is accessible to the reader and reasonably smooth. Propositions are usually justified and elaborated where necessary. The writer has produced a coherent text. Displays audience awareness and appropriate register	Grammatical structures are mostly or always accurate. In isolated instances, non-idiomatic or otherwise inappropriate expressions may occur but communication is never impeded. There is access to a range of structures. More complex structures are successfully attempted	None of the content demanded of the writer is omitted. The writer displays sensitivity to the text and to the audience. Task completed fully	Incorporates some additional vocabulary. Accurately corrects grammatical errors. Uses appropriate grammatical forms and accurately switches between grammatical forms. Adds appropriate cohesive elements. May accurately use colloquial language	Uses grammatically correct opening paragraph with an attempt to orientate or give background information to the reader. Uses headings where appropriate. Uses a style and tone (register) appropriate for purpose of writing. Includes all facts/information from the original. Introduces new paragraph(s) where appropriate; completes task, appropriate correct ending	
At the BM	Propositions are not always justified, elaborated or illustrated with examples where necessary although the links between one proposition and another are logical. The writer displays some sensitivity to the text and to the audience. The text is almost always coherent	Grammatical structures are generally accurate but errors may occasionally occur when more complex structures are attempted. Comprehension is seldom impeded. Some complex structures are attempted	Most of the content demanded of the writer is contained in the text. The writer displays some sensitivity to the text and to the audience. Task completed	Accurately corrects most grammatical errors. Combines ideas in sentences effectively. Uses appropriate grammatical forms and switches between grammatical forms are mostly accurate. Adds some cohesive elements (there may be some misuse or overuse). Seldom omits words	Attempts to improve opening paragraph to orientate or give background information to the reader; includes the majority of the facts/information from the original; introduces new paragraph(s) where appropriate. Shows evidence of improving style and tone (register) of writing. Completes task	

(continued)

(continued)

		Task 1: Expository writing		Task 3: Rewriting student composition		
		Scale 1: Organisation and coherence	Scale 2: Grammatical accuracy	Scale 3: Task completion	Scale 4: Vocabulary and grammar	Scale 5: Organisation and presentation of facts/information
Below the BM	Propositions may stand alone, without justification and without logical links. The text is difficult to follow and coherence may be flawed. Limited awareness of the audience is displayed	Grammatical errors occur regularly and may sometimes impede the reader's understanding. Few complex structures are attempted	Some of the content demanded of the writer may be omitted. Limited awareness of the audience is displayed hindering the full completion of the task	Sometimes fails to correct grammatical errors accurately. May fail to correct spelling errors. Sometimes omits words or adds inappropriate words. Often fails to use prepositions correctly. Errors in subject/verb agreement. Uses a controlled or restricted range of language to maintain accuracy and correctness of rewrite. Fails to add cohesive elements which enable the text to flow	Opening paragraph is not improved where required. Omits a few facts/information. May fail to introduce new paragraph(s). Rewrite/presentation of facts and information may be inaccurate or could be more fully developed but task is generally completed. May fail to or inadequately punctuate text, inappropriate ending	
Well below the BM	Propositions stand alone, without justification and without logical links. The text is virtually impossible to follow and is incoherent. Little awareness of the audience is displayed	Most of the text contains grammatical errors, causing comprehension to break down completely at times. Access to basic structures is clearly inadequate, and communication with the reader is often impeded	Much of the content demanded of the writer may be omitted. Little awareness of the audience is displayed. Task not completed	Fails to correct most grammatical errors accurately (e.g. subject/verb agreement, tense). Fails to correct spelling or confusion with he/she. Frequently omits words. Repeats phrases or linguistic elements. May misuse vocabulary and/or idioms. May have inadequate or missing cohesive elements	Opening paragraph is not improved where required. Omits or misrepresents facts/information. Rewrite is not developed in any way or is developed inappropriately. May fail to complete task in a satisfactory way	

Appendix E: Speaking Test Scales and Descriptors

Task 1A & 1B: reading aloud		Task 1C: Retelling a story/an experience/presenting arguments		Task 2: Professional oral interaction		Task 6: Explaining language matters	
Scale 1: Pron th , stress and intonation	Scale 2: Reading aloud	Scale 3: Grammatical accuracy	Scale 4: Organisi th n and cohesion	Scale 5: Interacting with peers	Scale 6: Explaining language matters	Scale 7: Organising discourse	Scale 8: Explaining language matters
5	<p>Pronunciation is completely error-free with no noticeable L1 characteristics, and candidate is very confident about the pronunciation of all words. Any mistakes that occur can be categorised as lapses rather than systematic errors. Sentence stress and intonation patterns are always appropriate and reading of the text is clear and effective as classroom communication</p>	<p>Speed of delivery and pausing are always appropriate. The speaker displays an extremely high degree of sensitivity to the text and to the audience and uses paralinguistic features effectively to communicate text</p>	<p>Grammatical structures are always accurate, with few or no occurrences of non-idiomatic or other inappropriate expressions. The candidate has access to a wide range of structures, which can be invoked at any time. Any 'mistakes' that occur can be categorised as lapses rather than systematic errors</p>	<p>A wide variety of appropriate means for connecting utterances are used. Relationships among concepts and/or ideas are clearly expressed, appropriately signalled and are never confused or confusing. The smooth and logical flow of ideas in the discourse facilitates the interlocutor's understanding. An extensive range of appropriate vocabulary is used</p>	<p>A very strong ability to talk easily, confidently and knowledgeably with peers in a professional manner about student language problems is evident. Full control over the conversational strategies of initiation, turn-taking, responding and disagreeing is evident, together with the ability to keep the discussion focused</p>	<p>The ability to organise discourse to explain student language problems to peers is clearly evident. Demonstrates control over and displays familiarity with a wide range of appropriate metalanguage without confusing peers. Able to produce appropriate examples to illustrate explanations. Explanations are fully coherent and easy to follow. Clarification and reformulation are never required because of the speaker's lack of ability</p>	<p>The ability to organise discourse to explain student language problems to peers is evident. Demonstrates control over a range of appropriate metalanguage. Able to produce appropriate examples to illustrate explanations. Explanations are coherent. Clarification and reformulation are seldom required because of the speaker's lack of ability</p>
4	<p>There may be isolated errors in the pronunciation of sounds and/or word-stress and a few L1 characteristics may be noticeable (e.g. minor problems with consonantal clusters). However, the candidate is confident about the pronunciation of words. Sentence stress and intonation patterns are appropriate and reading of the text is more than acceptable for classroom communication</p>	<p>In general, speed of delivery and pausing are optimally helpful. The speaker displays a high degree of sensitivity to the text and to the audience</p>	<p>Grammatical structures are mostly or always accurate. In isolated instances, non-idiomatic or otherwise inappropriate expressions may occur but communication is never impeded. There is access to a range of structures. More complex structures are successfully attempted</p>	<p>A variety of appropriate means for connecting utterances are used. Relationships among concepts and/or ideas are clearly expressed and hardly ever confused. There is a coherent and logical flow of ideas in the discourse. A good range of appropriate vocabulary is incorporated into task</p>	<p>The ability to talk easily, confidently and knowledgeably with peers in a professional manner about student language problems is evident. Control over the conversational strategies of initiation, turn-taking, responding and disagreeing is displayed, together with the ability to keep the discussion focused</p>	<p>The ability to organise discourse to explain student language problems to peers is evident. Demonstrates control over a range of appropriate metalanguage. Able to produce appropriate examples to illustrate explanations. Explanations are coherent. Clarification and reformulation are seldom required because of the speaker's lack of ability</p>	<p>The ability to organise discourse to explain student language problems to peers is evident. Demonstrates control over a range of appropriate metalanguage. Able to produce appropriate examples to illustrate explanations. Explanations are coherent. Clarification and reformulation are seldom required because of the speaker's lack of ability</p>

(continued)

(continued)

Task 1A & 1B: reading aloud		Task 1C: Retelling a story/an experience/presenting arguments		Task 2: Professional oral interaction		
Scale 1: Prof ⁱⁿ , stress and intonation	Scale 2: Reading aloud	Scale 3: Grammatical accuracy	Scale 4: Organi ⁱⁿ and cohesion	Scale 5: Interacting with peers	Scale 6: Explaining language matters	
3	<p>Although there may be some errors in the pronunciation of sounds and/or word-stress and a number of L1 characteristics are evident, pronunciation is unlikely to present comprehension problems for L2 learners. The candidate is fairly confident about the pronunciation of words. Sentence stress and intonation patterns may sometimes be inappropriate but reading of the text is seldom impeded and is acceptable for classroom communication</p>	<p>In part, features such as speed and pausing indicate a fairly high level of audience awareness. However, the speaker may occasionally read the text in a manner inappropriate to the text and may occasionally lose sight of the audience but, despite occasional lapses, manages to communicate meaning adequately</p>	<p>Grammatical structures are generally accurate but errors may occasionally occur when more complex structures are attempted. Sometimes, the teacher may recognise these errors and self-correct. Some reformulation is attempted. Some complex structures are attempted. Communication is seldom impeded</p>	<p>Connections among utterances are usually marked although there may be occasional confusing relationships among ideas. Ideas are generally presented logically with few, if any, examples of incoherent discourse</p>	<p>An ability to interact with peers in a discussion about student language problems is evident. Control over most of the conversational strategies involved in competently participating in a discussion, including an ability to keep the discussion focused, is displayed</p>	<p>Some ability to organise discourse to explain student language problems to peers using appropriate metalanguage is evident. Can usually produce appropriate examples to illustrate explanations. Explanations are mainly coherent, but there are occasional errors which may provoke requests for clarification or reformulation</p>
2	<p>The speaker displays some awareness of audience though may find it difficult to attend to the audience as well as the text throughout the reading. Speed and/or style of delivery are often inappropriate and may be hesitant. Meaning is not always communicated adequately</p>	<p>Grammatical errors (including subject/verb agreement, distinguishing between countable and uncountable nouns) occur in many utterances and are frequently abusive for the listener. Sometimes errors may be monitored and corrected. Few complex structures are attempted. Little attempt is made at reformulation</p>	<p>The connections between utterances are not adequately marked. There may be frequent confusion of relationships among ideas. Ideas are sometimes presented in an illogical manner with little focus on or relevance to the topic being discussed. A limited range of vocabulary is used</p>	<p>A limited ability to interact with peers in a discussion of student language problems is evident. Attempts to interact with the group during the discussion are occasional and limited. Only a limited ability to employ appropriate conversational strategies to keep the discussion focused is demonstrated</p>	<p>Limited ability to organise discourse to explain student language problems to peers is evident. Demonstrates lack of coherence. An inability to use appropriate metalanguage for explanations may hinder full understanding by peers. Displays a limited ability to produce appropriate examples. Clarification or reformulation are often necessary because of the speaker's lack of ability</p>	
1	<p>Frequent errors in the pronunciation of sounds, stress and intonation make communication difficult and lead to frequent interference with communication</p>	<p>The speaker's attention is wholly taken up by the effort of reading the text, and there is little evidence of audience awareness. There may be failure to sustain sense groups⁸ and meaning may often fail to be communicated</p>	<p>Most utterances contain grammatical errors, causing comprehension to break down completely at times. Access to basic structures is clearly inadequate, and communication is often impeded. No attempts are made at reformulation</p>	<p>Disjointed utterances are produced. There is no logical flow of ideas. The discourse is often incoherent. Communication breaks down frequently, often requiring prompting from the interlocutor</p>	<p>An extremely limited ability to interact with peers in a discussion of student language problems is evident. The speaker is unable to demonstrate competence in the conversational strategies required to interact professionally with peers</p>	<p>The teacher demonstrates little or no ability to explain common student language problems to peers and lacks the ability to use appropriate metalanguage and examples in such explanations. Explanations are almost impossible to follow</p>

Appendix F: CLA Scales and Descriptors (from Consultancy Study)

	The language of presentation/practice	Interaction with students	Grammatical accuracy	Pronunciation
4 Complete ability	There is evidence of a very strong ability to use English in the formal section of the lesson when presentation and practice take place. The ability to organise discourse and use appropriate cohesive and other signalling devices in order to alert the students to the various stages of the presentation is clearly evident. Explanations are always clear and coherent	Interaction with students demonstrates a very high level of sensitivity to student responses, an ability to always react appropriately to student initiation. Demonstrates an ability to hear and react to responses even when they are incomplete or lacking in coherence. There is no evidence of teacher language problems which can impede communication with students	Grammatical structures are always accurate, with no occurrence whatsoever of non-idiomatic or other inappropriate expressions. There is access to a wide range of structures, and these can be invoked at any time. Any 'mistakes' that occur can be categorised as lapses rather than systematic errors	Pronunciation is completely error-free with no noticeable L1 characteristics. Any mistakes that occur can be categorised as lapses rather than systematic errors. Sentence stress and intonation patterns are always appropriate, and communication is never impeded in the slightest
3 Extensive ability	The ability to use English in the formal section of the lesson when presentation and practice is demonstrated. There is evidence of the ability to organise discourse and to use appropriate cohesive and other signalling devices in order to alert the students to the various stages of the presentation. Explanations are usually clear and coherent	Interaction with students is usually smooth and natural whether on an individual or group basis. Demonstrates the ability to elicit, question, initiate and respond appropriately in order to foster communication with students. Communication with students is hardly ever impeded by teacher language problems. Demonstrates the ability to interact appropriately even when student responses are inaccurate or inappropriate	Grammatical structures are mostly or always accurate. In isolated instances, non-idiomatic or otherwise inappropriate expressions may occur but communication is never impeded. There is access to a range of structures. More complex structures are successfully attempted	There may be isolated errors in the pronunciation of sounds and/or word-stress and a few L1 characteristics may be noticeable (e.g. minor problems with consonantal clusters) Sentence stress and intonation patterns are appropriate. Communication is never impeded

(continued)

(continued)	The language of presentation/practice	Interaction with students	Grammatical accuracy	Pronunciation
2 Moderate ability	The ability to use English in the formal section of the lesson when presentation and practice take place is demonstrated. Able, with occasional errors, to organise discourse and use appropriate cohesive and other signalling devices in order to alert the students to the various stages of the presentation. Explanations are usually clear and coherent with little need for re-explanation or representation because of the speaker's lack of ability	Interaction with students is generally smooth and natural, whether on an individual or group basis. Demonstrates the ability to elicit, question, initiate and respond appropriately in order to foster communication with students. Communication with students is sometimes, but not often, impeded by teacher language problems. Demonstrates the ability to interact appropriately even when student responses are inaccurate or inappropriate	Grammatical structures are generally accurate, but errors may occasionally occur when more complex structures are attempted. Communication is seldom impeded. Some reformulation is attempted	Pronunciation of sounds is generally acceptable although there are some errors in the pronunciation of sounds and/or word-stress and a number of L1 characteristics are evident but are not obtrusive. Sentence stress and intonation patterns may sometimes be inappropriate, but communication is seldom impeded
1 Limited ability	Problems are encountered when using English in the formal section of the lesson when presentation and practice takes place. The ability to organise discourse and use appropriate cohesive and other signalling devices in order to alert the students to the various stages of the presentation is sometimes lacking. Explanations are sometimes unclear and lacking in coherence. Re-explanation or representation is often necessary because of the speaker's lack of ability	During interaction with students on an individual or group basis, barely able to communicate and/or encounters serious problems in communicating effectively with students. The ability to initiate and interact with or provide appropriate feedback to students is often lacking	Grammatical errors occur in some utterances and sometimes impede communication. Sometimes, the teacher may recognise these errors and self-correct. Few complex structures are attempted. Little attempt is made at reformulation	Pronunciation of sounds is generally acceptable although there are a number of errors in the pronunciation of sounds and/or word-stress and many L1 characteristics (e.g. consonantal clusters—'p/br', 'l/nr', 'v/w', 'th/f' problems) are obtrusive. The listener may experience some strain understanding the speaker, and communication is occasionally impeded
0 No ability	Serious problems occur when using English in the formal section of the lesson when presentation and practice take place. There is little evidence of the ability to organise discourse and use appropriate cohesive and other signalling devices in order to alert students to the various stages of the presentation. Explanations are unclear and lack coherence	During interaction with students on an individual or group basis, displays little or no ability to communicate effectively with students	Most utterances contain grammatical errors, causing comprehension to break down completely at times. Access to basic structures is clearly inadequate and communication is often impeded. No attempts are made at reformulation	Frequent errors in the pronunciation of sounds, stress and intonation make communication difficult and lead to frequent interference with communication

Appendix G: CLA Assessment Scales and Descriptors (Finalised)

	The language of instruction	The language of interaction	Grammatical accuracy	Pronunciation, stress and intonation
5	When required, the teacher demonstrates very strong ability to use English as the language of presentation. The teacher's ability to organise discourse and use appropriate signalling devices in order to alert students to the various stages of a presentation is clearly evident. Classroom instructions are invariably clear, comprehensible and appropriate for the level of the class	During interaction with students, the teacher demonstrates a very high level of linguistic awareness and sensitivity to student responses and an ability to always react in an appropriate linguistic manner to student initiation, such as a query, question or request for clarification, whenever it is required. The teacher demonstrates the language ability to be aware of and react to student responses even if these are incomplete or lacking in coherence. There is no evidence of teacher language problems that can impede communication with students	Grammatical structures are almost invariably accurate, with extremely limited, if any, occurrences of inappropriate expressions. Any 'mistakes' that occur can be categorised as 'slips' rather than systematic errors	Pronunciation is completely error-free with no noticeable first language (L1) characteristics. Any mistakes that occur can be categorised as 'slips' rather than systematic errors. Sentence stress and intonation patterns are always appropriate, and communication is never impeded in the slightest
4	When required, the teacher demonstrates good ability to use English in the language of presentation. There is evidence of an ability to organise discourse and use appropriate signalling devices in order to alert students to the various stages of the presentation. Explanations are almost always clear and coherent. Classroom instructions are almost always clear and understandable	The language of interaction with students is smooth and natural whether on an individual, group or whole-class basis. Whenever it is required, the teacher uses appropriate language to elicit, question, initiate and respond appropriately in order to foster communication with students. Repetition is rarely required because of student problems with teacher discourse. Communication with students is hardly ever impeded by teacher language problems. The teacher uses appropriate language to interact with students even when student responses are inaccurate or inappropriate	Grammatical structures are mostly accurate. In isolated instances, inappropriate expressions may occur but communication is not impeded	There may be isolated errors in the pronunciation of sounds and/or word-stress and a few L1 characteristics may be noticeable. Sentence stress and intonation patterns are appropriate. Communication is never impeded

(continued)

(continued)	The language of instruction	The language of interaction	Grammatical accuracy	Pronunciation, stress and intonation
3	<p>When required, the ability to use English adequately is demonstrated in the language of presentation. The teacher is able, with occasional 'slips' and errors, to organise discourse and use appropriate signalling devices in order to alert students to the various stages of the presentation. Explanations and classroom instructions are usually clear but will occasionally require re-explanation or representation because of problems with the speaker's language ability to organise the discourse and present it without ambiguity or confusion</p>	<p>The language of interaction with students is generally smooth and natural, whether on an individual, group or whole-class basis. Whenever it is required, the teacher generally uses appropriate language to elicit, question, initiate and respond appropriately in order to foster communication with students, even though language errors by the teacher or misunderstanding by students may occasionally impede interaction. When teacher language errors or student misunderstanding occur, a new phase of interaction, with appropriate adjustment to the previous utterances, will be initiated. There is evidence of generally appropriate use of language to acknowledge student responses. The teacher generally uses appropriate language to interact with students even when student responses are inaccurate or inappropriate</p>	<p>Grammatical structures are generally accurate, but errors may occasionally occur if more complex structures are attempted. Sometimes the teacher may recognise these errors and self-correct. Communication is seldom impeded. Some reformulation is attempted</p>	<p>Pronunciation of sounds is generally acceptable although there are some errors in the pronunciation of sounds and/or word-stress, and a number of L1 characteristics are evident but are not obtrusive. Sentence stress and intonation patterns may sometimes be inappropriate but communication is seldom impeded</p>
2	<p>A number of problems are encountered when using English if the presentation of a point of language is required or when giving instructions to students. The ability to organise discourse and use appropriate signalling devices in order to alert students to the various stages of the presentation is often lacking. Explanations and instructions are sometimes unclear and/or confused and confusing. Re-explanation or representation is often necessary</p>	<p>During interaction with students on an individual, group or whole-class basis, the teacher is barely or rarely able to use language appropriately with students and/or encounters serious problems in communicating effectively with students. The language of interaction with the whole class, group or individuals is often inappropriate for the level of student proficiency either because of problems of coherence or register. The language for classroom instructions is inappropriate and/or is not usually amended. The ability to use appropriate language to initiate and interact with or provide appropriate feedback to students is often lacking</p>	<p>Grammatical errors occur consistently in many utterances and sometimes may impede communication. The teacher consistently makes classic errors, fails to recognise or self-monitor such errors and thus fails to correct them. There are only rare examples of monitoring and self-correction of other errors. Few complex structures are attempted. Little attempt is made at reformulation</p>	<p>Pronunciation of sounds is almost acceptable although there are a number of significant errors in the pronunciation of sounds and/or word-stress, and many L1 characteristics are obtrusive. The student listener may experience strain or difficulty in understanding what the teacher says because of teacher pronunciation, stress or intonation errors, and communication is occasionally impeded</p>

(continued)

(continued)

	The language of instruction	The language of interaction	Grammatical accuracy	Pronunciation, stress and intonation
1	<p>Very serious and regular problems occur if the teacher uses English while presenting a point of language or giving instructions to students. There is little evidence of an ability to organise discourse and use appropriate signalling devices in order to alert students to the various stages of the presentation. Explanations and instructions are unclear and lack coherence</p> <p>Insufficient data on which to make an assessment</p>	<p>During interaction with students on an individual, group or whole-class basis, the teacher displays little or no ability to use language appropriately to interact effectively with students</p>	<p>Most utterances contain grammatical errors, causing comprehension to break down completely at times. Access to basic structures is clearly inadequate and communication is often impeded. Self-monitoring and self-correction never occur. No attempts are made at reformulation</p>	<p>Frequent errors in the pronunciation of sounds, stress and intonation make communication difficult and lead to frequent interference with communication</p>
0	<p>Insufficient data on which to make an assessment</p>	<p>Insufficient data on which to make an assessment</p>	<p>Insufficient data on which to make an assessment</p>	<p>Insufficient data on which to make an assessment</p>

References

- Andrews, S. J. (1999). *The metalinguistic awareness of Hong Kong secondary school teachers of English* (Unpublished Ph.D. thesis). University of Southampton.
- Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.
- Carlson, R. E. (1986). *Field-test data vs. real-test data*. Paper presented at the 67th Annual Meeting of the American Educational Research Association, San Francisco, CA.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Coniam, D. (1995). Towards a common ability scale for Hong Kong English secondary school forms. *Language Testing*, 12(2), 182–193.
- Coniam, D., & Falvey, P. (1996). *Setting language benchmarks for English language teachers in Hong Kong secondary schools*. Hong Kong: Advisory Committee on Teacher Education and Qualifications.
- Coniam, D., & Falvey, P. (1998). *Validating the classroom language assessment component: The Hong Kong English language benchmarking initiative*. Hong Kong: Advisory Committee on Teacher Education and Qualifications.
- Coniam, D., & Falvey, P. (1999). The English language benchmarking initiative: A validation study of the classroom language assessment component. *Asia Pacific Journal of Language in Education*, 2(2), 1–35.
- McNamara, T. (1996). *Measuring second language performance*. New York: Longman.

David Coniam is Chair Professor and Head of Department of the Department of Curriculum and Instruction in the Faculty of Education and Human Development at The Education University of Hong Kong, where he is a teacher educator, working with teachers in Hong Kong primary and secondary schools. His main publication and research interests are in language assessment, language teaching methodology and computer assisted language learning.

Peter Falvey is a teacher educator, formerly a Head of Department in the Faculty of Education, The University of Hong Kong. His main publication and research interests are in language assessment, first and second language writing methodology, and text linguistics.

Chapter 6

The English Language Benchmark Subject Committee



David Coniam and Peter Falvey

Abstract This chapter details the work of the English Language Benchmark Subject Committee (ELBSC) in developing, moderating, amending, changing and overseeing further the work on language benchmark developments.

The English Language Benchmark Subject Committee—Purpose and Brief

The English Language Benchmark Subject Committee (ELBSC) was convened in October 1997 under the auspices of the then Hong Kong Examinations Authority (HKEA) which later became the Hong Kong Examinations and Assessment Authority (HKEAA). Its purpose was to produce language benchmark specifications and an assessment syllabus for promulgation to Hong Kong teachers of English language in preparation for a large-scale pilot exercise—the Pilot Benchmark Assessment (English) (PBAE). The objective of the PBAE was to examine the prototype benchmark tests which the ELBSC had recommended, and to trial these tests on as representative a sample as possible of the Hong Kong English language teacher cohort. The composition of the ELBSC was very broad. The time frame the ELBSC was given was one year, using the consultancy report (Coniam & Falvey, 1996) as the starting point for the ELBSC's initial discussions. There was considerable debate over the substance of the report. While the majority of the recommendations were accepted by the ELBSC—that is the areas to be assessed—certain details of how assessment might be accomplished—the format of the Reading and Listening Tests,

D. Coniam (✉) · P. Falvey
Department of Curriculum and Instruction, Faculty of Education & Human Development,
The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, Hong Kong
e-mail: coniam@eduhk.hk

P. Falvey
e-mail: falvey@eduhk.hk

for example, and the scales and descriptors of the CLA, were not wholly accepted by the ELBSC. To resolve questions that the ELBSC raised, five Working Party sub-groups were formed under the ELBSC, each tasked with investigating one of the five areas to be assessed, namely Reading, Writing, Speaking, Listening and Classroom Language Assessment.

Pre-PBAE Validation Studies

For some of the test types, validation exercises of the test material, or of the training and standardisation of assessors were conducted by the consultants, with subjects consisting of in-service and pre-service teachers at local universities. Eight reports were produced by the consultants focusing on the validation of the assessment instruments and the training and standardisation of assessors for the criterion-referenced tests.

The reports contain detailed information on different aspects of the development of the English Language Benchmarking Initiative. These reports were:

1. Validating the Classroom Language Assessment Component: The Hong Kong English Language Benchmarking Initiative (Coniam & Falvey, 1998a)
2. Validating the Reading Test: The Hong Kong English Language Benchmarking Initiative (Coniam & Falvey, 1998b)
3. Piloting the Multiple-Choice Cloze Test: The Hong Kong English Language Benchmarking Initiative (Coniam & Falvey, 1998c)
4. Validating the Speaking Test: The Hong Kong English Language Benchmarking Initiative (Coniam & Falvey, 1998d)
5. Pre-pilot Exercise Rewriting and Speaking Components of the English Language Benchmark Project (Falvey & Coniam, 1998a)
6. Assessor Training and Standardisation for Classroom Language Assessment: The Hong Kong English Language Benchmarking Initiative (Falvey & Coniam, 1998b)
7. Assessor Training and Standardisation for the Speaking Test: The Hong Kong English Language Benchmarking Initiative (Falvey & Coniam, 1999c)
8. Assessor Training and Standardisation for the Writing Test: The Hong Kong English Language Benchmarking Initiative (Falvey & Coniam, 1999d).

The studies will be referred to from time to time in this and the following chapters.

As the list above reveals, no trialling of any material for the Listening Test was possible. The Listening Test that emerged suffered, not surprisingly, from the deficiencies that are discussed further in this chapter below.

The Work of the ELBSC

Between October and December 1997, the ELBSC met 32 times. The discussion and recommendations made by the ELBSC for test types are now described. A considerable number of amendments and changes—as might be expected—were made to the original recommendations of the 1996 consultancy feasibility study report by the ELBSC as a result of their deliberations.

Classroom Language Assessment

The CLA was discussed at length in the ELBSC because it would be a performance-based test that would take place in a live taught class. While the ELBSC was very much in agreement with the philosophy behind the use of an authentic test, logistic concerns were expressed at the administration of a live CLA.

Although English language teachers are used to paper-and-pencil tests, a live classroom test would be much more threatening. The constructs assessed would need to be broad in terms of language skills that were assessed, i.e. that they should not be biased against any particular group—primary versus secondary, for example. Care also had to be taken that the constructs which were to be established involved the assessment of language only and not pedagogical skills or personality traits. Support for the retention of CLA was made in a 1999 Colloquium on English Language Benchmarks held in Hong Kong, where Nevo (1999) stated unequivocally that the inclusion of the CLA in language benchmarking should be retained in spite of inevitable arguments that it would be costly and time-consuming.

A Working Party for CLA was formed under the main ELBSC to examine the constructs that the consultancy team had formulated in their original 1996 report and to examine the constructs, scales and descriptors both for validity and potential reliability. The Working Party met six times, watched over 20 videos, discussed the skills and constructs they felt appropriate to English language teachers, and reported back to the ELBSC.

There was strong agreement that the four constructs that had been formulated in the Consultancy Report for English Language Benchmarks (Coniam & Falvey, 1996) were the essential English language skills which teachers of English language required in order to underpin the effective teaching of English. *Grammatical Accuracy* and *Pronunciation, Stress and Intonation* are the two ‘formal’ elements which define an English language teacher’s ability in English. The other two elements *The Language of Presentation/Practice* and *The Language of Interaction* are the functional realisations of a teacher’s formal ability in English in terms of communicating with students and getting things done in the classroom. Scales and their descriptors were then formed to reflect those skills at various levels of ability.

The four constructs and their associated descriptors of language performance were arrived at by the following methods:

1. Observation of English language lessons on video,
2. Creation of a taxonomy of teacher language tasks,
3. Development of prototype constructs,
4. Moderation of the constructs by experts and practising teachers,
5. Creation of scales,
6. Creation of descriptors for each scale based upon distinct levels of language performance,
7. Validation of the constructs and descriptors through moderation and empirical study and
8. Submission of the prototypes to the ELBSC.

After phase (7), Level '3' of the prototype scales was adopted as the tentative benchmark level. A new Level '0' was added to indicate that no performance in that skill area was available for grading, e.g. speaking in Cantonese for the whole lesson. By mid-1998, the specifications of the scales after revision, modification and amendment were resolved as follows:

1. Grammatical Accuracy,
2. Pronunciation, Stress and Intonation,
3. The Language of Interaction and
4. The Language of Instruction.

The Speaking Test

Specifications

As reported in the validation study of the Speaking Test (Falvey & Coniam, 1999), the ELBSC agreed, after reviewing the different options proposed in the Consultants' Report for English Language Benchmarking (Coniam & Falvey, 1996), that the assessment of speaking was a crucial part of the English language benchmark assessment procedure.

In addition, the ELBSC eventually decided that some skills must be assessed for all teachers of English language, e.g. the comparatively difficult and teacher-specific skill of reading aloud; and the language teacher skill of storytelling or recounting. The 1996 Consultancy Report proposed three test types and seven separate scales for the Speaking Test.

Although the ELBSC'S deliberations on the Speaking Test retained the essence of the consultants' 1996 recommendations in that the test still consisted of three linked elements, certain elements and task types were changed.

As can be seen from Table 6.1, one of the original scales (Pronunciation, Stress and Intonation) was tested twice in the original proposals so one of those pronunciation scales was dropped. The test types and scales which the ELBSC accepted and on

Table 6.1 Scales and descriptors in the 1996 Consultancy Report

Test type	Scale	Salient linguistic features
1. Reading aloud: giving instructions	1. Pronunciation, Stress and Intonation	Sounds, stress, intonation
	2. Reading Aloud with Meaning	Speed of delivery, pausing, awareness of audience
2. Talking about teaching	1. Pronunciation, Stress and Intonation	Sounds, stress, intonation
	2. Grammatical Accuracy	Grammatical accuracy, range of structures
	3. Organisation and cohesion	Coherence, logical flow of ideas, relationships between ideas
3. Oral interaction	1. Interacting with Peers	Including turn-taking, initiating, responding, agreeing and disagreeing
	2. Explaining Language Matters to Peers	Including the use of appropriate metalanguage, appropriate examples

Table 6.2 Scales and descriptors proposed by the ELBSC

Test type	Scale	Salient linguistic features
1. Reading aloud a text	1. Pronunciation, Stress and Intonation	Sounds, stress, intonation
	2. Reading Aloud with Meaning	Speed of delivery, pausing, awareness of audience
2. Telling a story/recounting a personal experience/presenting arguments	1. Grammatical Accuracy	Grammatical accuracy, range of structures
	2. Organisation and Cohesion	Coherence, logical flow of ideas, relationships between ideas
3. Professional oral interaction	1. Interacting with Peers	Including turn-taking, initiating, responding, agreeing and disagreeing
	2. Explaining Language Matters to Peers	Including the use of appropriate metalanguage, appropriate examples

which the PBAE Speaking Test was based and which are reported here are presented in Table 6.2.

The six scales and the descriptors that were used in the PBAE are contained in Appendix E “[Speaking Test Scales and Descriptors](#)”, p. 81–85 Chap. 5.

Assessor Training for the PBAE

An investigation was conducted into the reliability of the Speaking Test assessors.

The purpose of the training was to train and standardise assessors. However, this also involved conducting an analysis of the assessors' scores in order to deselect potentially weak assessors, those who might be too harsh or too lenient or those unable to apply the scales and descriptors consistently.

The 16 assessors were first given the marking schemes and rating scales and given time to read and digest them. They were then shown the first set of three applicants and asked to rate them without discussion. The purpose of this blind rating was to enable the consultants to observe how much initial variability there was amongst the assessors. Subsequently, this variability was to be compared with their performance on the final ratings carried out at the end of the day.

After the first, blind, rating session, the trainee assessors were then given detailed training for three more full sessions and standardisation feedback and follow-up after they had given their grades. While there was an initial wide range of marks, this was reduced through the day's training to a much narrower range. Assessor-to-model misfit was also substantially reduced. A detailed description of the training and standardisation procedure for the Speaking Test is provided in Falvey and Coniam (2000)

All assessors remarked that they felt the assessor training session had been remarkably well organised and that they had benefited in terms of being prepared for assessing teachers on the PBAE Speaking Test. Many of the assessors' suggestions were adopted and incorporated into the PBAE Speaking Test. Examples of changes include the addition of a poem to the prose passage in the Reading Aloud section.

The Writing Test

The ELBSC agreed that the original construct that had been formulated in the Consultancy Report for English Language Benchmarks (Coniam & Falvey, 1996) was an essential facet of the English language skills which teachers of English language require in order to underpin the effective teaching of English.

The ELBSC retained the Expository Writing test type reported in this chapter but increased the number of levels from five to six by making Level 0 a description of 'no performance upon which to make an assessment'. The ELBSC also introduced a new test type (Rewriting) with the result that the Writing Test that was used in the PBAE finally consisted of two test types and five scales (drawn from the original test type and the new test type).

The new test type was an innovative test of writing awareness and writing skill. It requires test takers to rewrite a student essay (typically a low-level Secondary 5 [Year 11] essay, written for the HKCEE examination). The purpose of the rewriting task was to demonstrate that test takers can not only understand the problems associated

Table 6.3 Scales and descriptors proposed by the ELBSC for the rewriting task

Test type	Scale	Salient linguistic features
1. Writing professionally	Grammatical Accuracy	Grammatical accuracy, range of structures
	Organisation and Coherence	Organisation of text, coherence
	Task Completion	All tasks requested in the stimulus must be completed
2. Rewriting a student composition	Vocabulary and Grammar	Grammatical accuracy, range of structures, appropriate lexical choice
	Organisation and Presentation of Facts/Information	Logical flow of ideas, relationships between ideas, retention of main facts/information from the original student text

with the writing of the composition but also that they have the requisite skills to allow them to rewrite it in an acceptable/exemplary manner.

The task was trialled and found to work well. Once preliminary descriptors had been established, a sample batch of rewrites was given to consultants and HKU language education specialists who, acting as assessors, were asked to read the rewritten text and then use the prototype descriptors to assign a benchmark level to each text. Although, at this stage, little training was given to assessors, the assessors reported favourably on their ability to operationalise the descriptors. Adjustments were made to the prototype descriptors based on feedback from these assessors during a pre-PBAE pilot in 1998. Concurrent validity for the rewriting task was high with the expository writing task ($r = .66, p < .001$). Concurrent validity was also high with the other skills (calibrated MC items— $r = .63, p < .001$); Speaking Test—($r = .89, p < .001$).

The scales and descriptors used in the PBAE Writing Test are shown in full in Appendix D “[Writing Test Scales and Descriptors](#)“, p. 76–80. A summary is presented in Table 6.3.

The Reading Test

Reading

It was agreed, as for the Speaking Test, that a basic principle should be that teachers taking the test must be treated as mature adults and that multiple-choice tasks which resembled school tests should be avoided as far as possible. Principles laid down by the ELBSC were as follows:

- 1 It should tap higher-level reading skills.
- 2 It should neither duplicate HKEA school tests such as the HKCEE or HKASLE examinations nor appear similar to them. (This was for purposes of credibility and face validity. Teacher informants made it clear that they did not want to see a battery of tests which appeared to resemble the tests for which many of them were preparing their students.)
- 3 It should, ideally, *not* be in a multiple-choice format. (This criterion was established because of the ELBSC's desire to promote the more modern paradigm of assessment which eschews large-scale multiple-choice testing.)
- 4 The material should be authentic.
- 5 Its topic content should be based on domains that English language teachers might encounter in their professional lives, i.e. English language teaching and language education.

Cloze

The ELBSC also agreed eventually that a multiple-choice element should form part of the test battery.

Although initially resistant to the inclusion of multiple-choice test items, the ELBSC finally agreed to the inclusion of a multiple-choice cloze test because of the reliability such a test might afford the HKEA as an anchor against the Reading Test. The ELBSC stated, however, that:

1. The items should be integrated into a text type such as a cloze passage and not consist of discrete point items.
2. Some of the items should test discourse-level skills.
3. The items should be properly pretested.

Following the principle of using authentic material, a number of cloze passages were prepared for pretesting with as little amendment to their initial state as possible. Setters and moderators attempted to make as few amendments as possible to the original in order to provide teachers with the type of text that they could encounter in their professional lives. Item types included grammar and vocabulary, as do most cloze tests. However, an effort was also made to include items that required test takers to take the discourse context into account (c.f., Deyes, 1984).

The Listening Test

It was agreed that the stimulus for the Listening Test should consist of an authentic discussion, based around English language teaching/educational themes. It was decided not to use a single speaker as this would closely represent the academic listening skills required in a formal lecture. Consequently, a Listening Test was developed in which answers would be of an open-ended format. One of the ELBSC's recommendations was that the Listening Test should be delivered in a video rather than an audio format—the latter being the format adopted by the HKASLE Year 13 Use of English examination at the time.

A number of formats were experimented with. One of the formats involved the production of questions based on major themes rather than linear questions which paralleled the videotaped discussion. Participants in the video were briefed on the topic and then asked to take a stance on it. Topics covered included the use of native English-speaking teachers in Hong Kong secondary schools, the medium of instruction in schools and the role of English in education.

Unfortunately, the innovative video approach to a new test format was not piloted, due to logistical problems. As a result, the final format of the Listening Test can be described as a hybrid between a test for English language teachers and a 'more demanding' HKASLE Year 13 Use of English Listening Test. The majority of the questions generally paralleled the text (as with the Use of English Listening Test). Some questions did not, however, which required test takers to take a broader perspective, and to draw on different sections of the taped discussion. In addition, there was an attempt to include questions which required answers that drew on more than factual recall—the general item types used with questions in the HKASLE UE Listening Test.

As stated above, the wishes of the ELBSC were unable to be followed because of practical constraints (e.g. finding enough test rooms with video facilities for large numbers of test takers) so the HKEA decided that it would not be possible to administer the test which had been prepared for use on video via video. Instead, the video was converted to an audio tape for the live test. As reported below, this led to problems, reported by test takers. See also the discussion in Coniam (2001) of the relative lack of advantage of using video over purely audio as a medium for conducting listening tests and a further discussion of this topic in the closing chapter, Chap. 18.

(Produced by the English Language Benchmark Subject Committee for the Pilot Benchmark Assessment (English))

Part 1: Assessment purpose / target group / objectives / language model
Part 2: Overall statement including a discussion of constructs
Part 3: Major components
Part 4: Task and question types
Part 5: Syllabus specifications (number of sections / papers / parts / suggested text lengths / timing etc.)

Fig. 6.1 Framework for Pilot Benchmark Assessment (English Language)

From the ELBSC and Its Working Parties to Its Moderation Committees

By the end of 1997, the ELBSC had agreed on the composition of the benchmark test battery, and a draft test blueprint was produced. This is now reproduced in Fig. 6.1.

Part 1: Assessment Purposes

<i>Purpose of assessment</i>	To establish minimum, acceptable levels of language ability for teachers of English in lower forms of secondary school
<i>Target group</i>	All teachers of English in lower forms of secondary school ^a
<i>Objectives: Major</i>	To establish minimum, acceptable levels of teacher language ability in order to underpin the effective teaching of English in lower secondary school classrooms
<i>Objectives: Specific</i>	To establish minimum, acceptable levels of competence in order to deliver the English language curriculum in the classroom in the specific language skill areas of: <ul style="list-style-type: none"> • Classroom language • Speaking • Listening • Reading • Writing
<i>Language model</i>	A functional model of language (Halliday, 1985) with reference for language testing to Bachman and Palmer’s 1996 model of language (organisational [grammatical and textual] and pragmatic [functional and sociolinguistic]) knowledge and strategic (metacognitive strategies) competence

^aThe language skills of some upper primary and upper secondary teachers of English will also be sampled for purposes of comparison with their lower secondary counterparts

Part 2: Overall Statement Including a Discussion of Constructs

Construct Statement

The sections below contain construct descriptions of the major areas to be benchmarked. It will be noted that some overlap occurs. The reason for this is that some important language skills are used in different but relevant contexts of use. All of these contexts of use are deemed important for the effective practice of English both professionally (e.g., with colleagues and specialists) as well as in the classroom. Therefore, for example, it will be seen that the assessment of pronunciation occurs in two contexts—in a reading aloud task and in the context of the classroom with students. Grammar, likewise is assessed both in written (Writing component) and multiple-choice cloze form (Reading component) and, in addition, in two different but relevant spoken forms (speaking to peers/superiors and speaking to students in a live classroom context)

	Classroom Language Assessment
<i>To be examined by</i>	Education Department Classroom Language Assessors
<i>Authenticity</i>	
<i>Areas to be benchmarked</i>	<p>CLASSROOM LANGUAGE in which minimum, acceptable levels of ability to communicate with students appropriately are assessed in the areas of:</p> <ul style="list-style-type: none"> • Grammatical Accuracy • Pronunciation, Stress and Intonation • The Language of interaction • The Language of Instruction

	‘Formal’ Assessment: Direct and Indirect
<i>To be examined by</i>	Hong Kong Examinations Authority
<i>Authenticity</i>	In all cases, authentic texts, or adaptations of authentic material will be used
<i>Areas to be benchmarked</i>	<p>SPEAKING in which competence is assessed, when interacting with educated native and non-native speakers, in the language skills areas of:</p> <ul style="list-style-type: none"> • Pronunciation, Stress and Intonation • Reading Aloud with Meaning • Grammatical Accuracy • Organisation and Coherence • Interacting with Peers • Explaining Language Matters to Peers
	<p>LISTENING in which competence is assessed by listening to and understanding educated native and non-native speakers of English in audio/video recordings. Possible text types would be discussions, debates, interviews, documentaries and current affairs programmes which discuss matters broadly related to education and professional language teaching. These might be drawn directly from the English language media in Hong Kong or developed from authentic interviews, discussions etc.</p>
	<p>READING in which competence in reading and understanding texts of an agreed appropriate nature and level within the context of professional language teaching is assessed (e.g., texts taken from journals such as Modern English Teacher, English Language Teaching Journal, Curriculum Forum, Practical English Teacher, as well as fiction and newspaper articles on relevant topics)</p> <p>VOCABULARY, GRAMMAR and DISCOURSE in which minimum, acceptable levels of vocabulary, grammar, discourse and textual knowledge are assessed in a cloze procedure</p>
	<p>WRITING in which competence is assessed by means of:</p> <ul style="list-style-type: none"> • a stand-alone expository writing task • rewriting/improving a student composition

Part 3: Major Components

	Classroom Language Assessment
<i>Areas to be assessed</i>	Assessment of teacher language skills in a normal classroom working environment
<i>Components/scales</i>	<p>Classroom Language Assessment</p> <p>To assess teachers’ ability to use English for classroom purposes in the following ways:</p> <ul style="list-style-type: none"> • Grammatically • With appropriate pronunciation, stress and intonation <p>in order to demonstrate the communicative language skills which involve:</p> <ul style="list-style-type: none"> • The Language of Interaction, i.e.: <ul style="list-style-type: none"> – Eliciting – Responding – Providing feedback – The language of classroom management, including: praising/advising/acknowledging • The Language of Instruction, i.e. <ul style="list-style-type: none"> – Presentation – Giving instructions – Signalling
	‘Formal’ Assessment: Direct and Indirect
<i>Areas to be assessed</i>	<ul style="list-style-type: none"> • Speaking • Listening • Reading, Vocabulary, Grammar and Discourse • Writing
<i>Components/scales</i>	<p>SPEAKING</p> <ul style="list-style-type: none"> • Pronunciation, stress and intonation • Reading aloud with meaning • Grammatical accuracy • Organisation and coherence • Interacting with peers • Explaining language matters to peers
	<p>LISTENING</p> <p>Open-ended responses to audio/video-recorded spoken discourse</p>
	<p>READING</p> <p>Open-ended responses to texts</p> <p>VOCABULARY, GRAMMAR and DISCOURSE COMPONENT</p> <p>Multiple-choice cloze</p>
	<p>WRITING</p> <ul style="list-style-type: none"> • Organisation and coherence • Grammatical accuracy • Task completion

Part 4: Task/Question Types

Classroom Language Assessment

A live lesson conducted with the teacher's normal time-tabled class which would include a display of the language skill areas which have been specified in previous sections

'Formal' Assessment: Direct and Indirect

SPEAKING

- An integrated "Reading and Speaking" task consisting of:
 - Reading aloud, e.g., narrative, instructions, poem etc., thematically linked to:
 - Telling a story/recounting a personal experience/presenting arguments based on a stimulus provided, e.g., written prompts, an incomplete short story, a set of pictures or the passage for reading aloud
- Discussing student language problems presented within the context of an authentic student composition

LISTENING

Open-ended listening tasks based on English language teaching situations or topics of general educational interest in the form of an exposition, e.g., lecture situation, dialogue or debate with the following types of question: factual; attitudinal; inferential; gist/summary

READING

- Open ended reading tasks based on a text or texts provided
- VOCABULARY, GRAMMAR and DISCOURSE COMPONENT
- Multiple-choice cloze
-

WRITING

- An argumentative/explanatory/instructional writing task related to the professional or practical work of an English language teacher OR a writing task related to a text on a relevant language teaching topic
 - Improving a student composition by identifying and solving lexico-grammatical and discourse problems
-

Part 5: Syllabus Specifications

Classroom Language Assessment

Paper IV: Classroom Language Assessment

Note: A minimum of 5 days' notice will normally be given by the assessor(s) to the teacher

Briefing: The teacher will brief the assessor(s) before the class takes place. The briefing will include information on the students' previous language learning and teacher language skills to be demonstrated

Time: 5–15 min, as required by the teacher

NOTE: This part is not assessed

Assessment: Live lesson:

The assessment will take place in a single period. The first 10 minutes of the single period will not be assessed. This non-assessed section of the lesson will allow the teacher, assessor(s) and students to get used to each other

‘Formal’ Assessment: Direct and Indirect			
Major elements of the benchmarking assessment including: <ul style="list-style-type: none"> • number of sections • number of parts • text lengths • timing 	PAPER-AND-PENCIL TESTS This assessment consists of two papers:		This assessment consists of two sections:
	Paper I Reading and Writing <ul style="list-style-type: none"> • Part 1: Multiple-choice cloze Time: 30 min Text: approximately 500 words Items: 20–30 • Part 2: Reading Time: 1 h Text(s): One text of 1500–2000 words or two texts of 750–1000 words each Questions: about 20 of various types • Part 3: Writing time: 1 h 15 min • Text: Stand-alone writing task: stimulus material will be given as input for the writing task, either using the reading passage in the reading comprehension component or a different text of 200–300 words. Text: Improving a student composition task—a text of about 200–300 words will be used 	Paper II Listening <ul style="list-style-type: none"> • Listening and responding to an audio/video recording(s) which is/are heard only once Time: 1 h Preparation time: 3–5 min to look at the question paper Time for listening and responding: 30 min Completion time: 10–15 min • ‘Text’: One segment of spoken discourse of approximately 25–30 min or two segments of approximately 10–15 min each Questions: about 20 of various types 	

As the HKSAR Government wished to press ahead with the specimen material and prepare for the PBAE (see below), in early 1998 four Moderation Committees were formed under the aegis of the Hong Kong Examinations Authority to set two sets of test material for the four paper-and-pencil tests. One set was to be released as specimen material to teachers; the other set was to be live pilot test material. A booklet of the prototype benchmark syllabus together with specimen material was published in September 1998 by ACTEQ in the syllabus document *Syllabus Specifications, Specimen Questions, Notes for Classroom Language Assessment*.

At the same time, i.e. September 1998, the Education Bureau began canvassing schools in an attempt to recruit teachers to participate in the Pilot Benchmark Assessment (English), which is discussed in Chap. 7.

Summary

This chapter has discussed the pre-PBAE validation process by the ELBSC on the different components of the test battery, namely Classroom Language Assessment, Speaking, Writing, Reading and Listening. The work of the ELBSC contributed to the development of the Framework for Pilot Benchmark Assessment (English Language), which states the purposes, format and the structure of the PBAE. Chapter 7 describes the Pilot Benchmark Assessment phase of the consultancy study.

References

- Coniam, D. (2001). The use of audio or video comprehension as an assessment instrument in the certification of English language teachers: A case study. *SYSTEM*, 29(2), 1–14.
- Coniam, D., & Falvey, P. (1996). *Setting language benchmarks for English language teachers in Hong Kong secondary schools*. Hong Kong: Advisory Committee on Teacher Education and Qualifications.
- Coniam, D., & Falvey, P. (1998a). *Validating the classroom language assessment component: The Hong Kong English language benchmarking initiative*. Hong Kong: Advisory Committee on Teacher Education and Qualifications.
- Coniam, D., & Falvey, P. (1998b). *Validating the reading test: The Hong Kong English language benchmarking initiative*. Hong Kong: Advisory Committee on Teacher Education and Qualifications.
- Coniam, D., & Falvey, P. (1998c). *Piloting the multiple-choice cloze test: The Hong Kong English language benchmarking initiative*. Hong Kong: Advisory Committee on Teacher Education and Qualifications.
- Coniam, D., & Falvey, P. (1998d). *Validating the speaking test: The Hong Kong English language benchmarking initiative*. Hong Kong: Advisory Committee on Teacher Education and Qualifications.
- Deyes, T. (1984). Towards an authentic 'discourse cloze'. *Applied Linguistics*, 5(2), 128–137.
- Falvey, P., & Coniam, D. (1998a). *Pre-pilot exercise for the rewriting and speaking components of the English language benchmark project*. Hong Kong: Hong Kong Examinations Authority.

- Falvey, P., & Coniam, D. (1998b). *Assessor training and standardisation for classroom language assessment: The Hong Kong English Language Benchmarking Initiative*. Hong Kong: Advisory Committee on Teacher Education and Qualifications.
- Falvey, P., & Coniam, D. (1999c). *Assessor training and standardisation for the speaking test: The Hong Kong English Language Benchmarking Initiative*. Hong Kong: Advisory Committee on Teacher Education and Qualifications.
- Falvey, P., & Coniam, D. (1999d). *Assessor training and standardisation for the writing test: The Hong Kong English Language Benchmarking Initiative*. Hong Kong: Advisory Committee on Teacher Education and Qualifications.
- Falvey, P., & Coniam, D. (2000). Establishing English language writing benchmarks for primary and secondary teachers of English language in HongKong. *HongKong Journal of Applied Linguistics*, 5(1), 128–159.
- Halliday, M. A. K. (1985). *An introduction to functional grammar*. Edward Arnold: London.
- Nevo, D. (1999). Paper presented at Language Benchmarks Colloquium. September 29, 1999. Hong Kong: Hong Kong Institute of Education.

David Coniam is Chair Professor and Head of Department of the Department of Curriculum and Instruction in the Faculty of Education and Human Development at The Education University of Hong Kong, where he is a teacher educator, working with teachers in Hong Kong primary and secondary schools. His main publication and research interests are in language assessment, language teaching methodology and computer assisted language learning.

Peter Falvey is a teacher educator, formerly a Head of Department in the Faculty of Education, The University of Hong Kong. His main publication and research interests are in language assessment, first and second language writing methodology, and text linguistics.

Chapter 7

The Pilot Benchmark Assessment (English)



David Coniam and Peter Falvey

Abstract This chapter covers the Pilot Benchmark Assessment (English) (PBAE)—the test bed for all the constructs, their benchmarks and the associated assessment instruments which were developed in the two and a half years prior to its administration. The background to the exercise and the formation of the sample are first discussed. After this, a brief analysis of each test component is provided. The chapter rounds off with a brief description of test takers' reactions to the tests (analysis of certain questions, as well as a digest of written comments, from the post-test questionnaire, together with a discussion of key issues from the qualitative interviews that were conducted with about 10% of the test-taking cohort).

Introduction

The PBAE ran from late 1998 to early 1999, lasting four months because each teacher was observed twice teaching a live English language class. A Benchmark Assessment Unit, constituted under the then Education Department and trained by the consultants, began to assess teachers for the CLA component of the PBAE in late 1998, finishing all the assessments by early 1999. It involved two visits to over 320 teachers in their own classrooms. The pen-and-paper tests were subsequently administered in early February 1999.

The PBAE Sample

The PBAE sample was initially constructed so that it would be representative of lower secondary teachers of English. After discussion with ED's Statistics Section, the total

D. Coniam (✉) · P. Falvey
Department of Curriculum and Instruction, Faculty of Education & Human Development,
The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, Hong Kong
e-mail: coniam@eduhk.hk

© Springer Nature Singapore Pte Ltd. 2018
D. Coniam and P. Falvey (eds.), *High-Stakes Testing*,
https://doi.org/10.1007/978-981-10-6358-9_7

Table 7.1 Desirable and actual sample sizes

Type	Qualifications	Recommended sample size	Actual number
1	Teacher Cert., Subject-trained, Primary	40	30
2	Teacher Cert., Non-subject-trained, Primary	40	29
3	Teacher Cert., Subject-trained, Secondary	55	53
4	Teacher Cert., Non-subject-trained, Secondary	30	12
5	Degree, Relevant, Secondary	35	33
6	Degree, Relevant, Professional training, Secondary	35	35
7	Degree, Non-relevant, Secondary	110	86
8	Degree, Non-relevant, Professional training, Secondary	55	49
	Total	400	327

number of teachers for the PBAE was recommended as 400. Based on this figure, the ELBSC recommended the use of stratified sampling so that the individual subgroups of the target population would be proportionately represented in the sample (Hatch & Lazaraton, 1991). Column 3 in Table 7.1 presents the desired sample of participating teachers.

The Education Department therefore wrote to all school principals in Hong Kong asking them to provide detailed information of the English language teachers in their schools in order to construct the actual sample. The final figure for teachers who agreed to participate in the PBAE is presented in Column 4 of Table 7.1.

The participation rate for each group (with the exception of Group 4) was higher than 75%.

Since the PBAE required a representative sample, the presence of self-selecting volunteering teachers would not be appropriate (Hatch & Lazaraton, 1991; Rosenthal & Rosnow, 1975).

The PBAE and the 1996 Survey: Comparative Data

The composition of the PBAE sample may be seen in perspective by comparing relevant demographic data using the 1996 Consultancy Study survey as the anchor since the 1996 survey obtained data from almost 75% of the practising cohort of English language teachers in Hong Kong.

Table 7.2 presents figures for the distributions of teachers across Government, Aided and Private schools.

Table 7.2 School type

	Govt	Aided	Private	Total
1996 survey	858 (9.2%)	8373 (90.2%)	53 (0.6%)	9284
1999 PBEA	18 (5.9%)	268 (88.2%)	18 (5.9%)	304

Table 7.3 Desirable and actual sample sizes

Type	Qualifications	Number (%) 1999 PBAE	Number (%)—1996 q'aire survey
1	Teacher Cert., Subject-trained, Primary	30 (9.1%)	1919 (27.6%)
2	Teacher Cert., Non-subject-trained, Primary	29 (8.9%)	1056 (15.2%)
3	Teacher Cert., Subject-trained, Secondary	53 (16.2%)	371 (5.3%)
4	Teacher Cert., Non-subject-trained, Secondary	12 (3.7%)	55 (0.8%)
5	Degree, Relevant, Secondary	33 (10.1%)	671 (9.6%)
6	Degree, Relevant, Professional training, Secondary	35 (10.7%)	845 (12.1%)
7	Degree, Non-relevant, Secondary	86 (26.3%)	784 (11.2%)
8	Degree, Non-relevant, Professional training, Secondary	49 (15.0%)	1255 (18.0%)
	Total	327	6956

The PBAE had a much heavier weighting of secondary school teachers than primary school teachers. As stated above, this matched the objective of the PBAE whose purpose was to examine how lower secondary teachers performed on the prototype tasks.

Table 7.3 is a summary of the teacher qualification, primary/secondary distribution of the members of the PBAE sample. The table presents a final comparison of the numbers of teachers who participated in the 1999 PBAE and the representation of the cell type against the total English language teaching cohort in Hong Kong.

The total number of teachers who initially took part in the PBAE was 327, although some dropped out as the PBAE progressed over the four-month period. Table 7.4 presents, by test type, a breakdown of teachers who completed the PBAE, i.e. those who took the majority of the tests in the battery.

Table 7.4 Number of teachers taking part in the different test types

Classroom language assessment	297
Reading (Reading comprehension and cloze)	299
Writing	299
Listening	298
Speaking	303
Multiple-choice items	297

Analyses of the Tests Administered in the Pilot Benchmark Assessment (English)

The test components are discussed under two headings:

1. *Criterion-referenced tests* (Classroom Language Assessment, Speaking Test, Writing Test)
2. *Analytically marked tests* (Reading Test, Listening Test, Multiple-choice Test)

The discussion and analysis within each section encompasses, where relevant, detail on:

- Analysis of assessor performance (assessor means, inter-rater correlations, Many-Facet Rasch Analysis)
- Analysis of test statistics (means, item analyses, reliability coefficient, standard error)
- Analysis of test taker performance.

The PBAE Classroom Language Assessment Component

Introduction

The English Language Subject Benchmark Committee (ELBSC) agreed with and accepted the four constructs, presented in the Consultancy Report for English Language Benchmarks (Coniam & Falvey, 1996):

- Grammatical Accuracy
- Pronunciation, Stress and Intonation
- The Language of Interaction
- The Language of Instruction.

A full description of the validation procedures used for testing the operationalisability of the scales and descriptors was reported in Coniam and Falvey (1998a, b, c, d). The four constructs, their six scales and their associated descriptors are contained Appendix G “CLA Assessment Scales and Descriptors (Finalised)” in Chap. 5.

A Benchmark Assessment Unit was set up under the ED in 1998 to undertake the Classroom Language Assessment. Training consisted of 30 h, prior to CLA visits, as well as ongoing standardisation during the visits consisting of:

- seven three-hour seminar-based sessions
- three school visits plus three-hour follow-up sessions
- 15 paired-up visits at the beginning of the assessment period (See Falvey & Coniam, 1999).

Analysis of Test Taker Performance

As an indication of the overall performance of the cohort, Table 7.5 presents the mean for each scale. (The level recommended by the ELBSC as the potential benchmark was a Level ‘3’.)

An analysis of performance on the different scales revealed that the majority of teachers’ scores on the different scales were clustered at Levels ‘3’ and ‘4’. 17–18% of teachers scored at Level ‘5’, indicating that there were some extremely able language users in the classroom. The scale on which fewest teachers failed to reach the tentative benchmark was *Pronunciation, Stress and Intonation*.

Table 7.6 presents the numbers of teachers who achieved the putative CLA benchmark.

Table 7.5 Means for the four CLA scales

Scale	<i>N</i>	Mean	SD	Min.	Max.
Grammatical accuracy	304	3.66	0.87	1.00	5.00
Pronunciation, stress and intonation	304	3.69	0.83	1.00	5.00
The language of instruction	301	3.66	0.87	1.00	5.00
The language of interaction	300	3.64	0.89	1.00	5.00

Table 7.6 Test takers achieving a 3 (or better)

No. of scales	No. of test takers
On all 4 scales	260 (84.4%)
On 3 scales	19 (6.2%)
On 2 scales	19 (6.2%)
On 1 scale	6 (1.9%)
On 0 scales	4 (1.2%)
Total	308

Table 7.6 reveals that 84.4% of the teachers observed achieved the prototype minimally acceptable benchmark grade of ‘3’ on all four scales. The ELBSC had striven to formulate what the desirable ‘minimum acceptable level of English language’ should be for an English language teacher. It would appear, therefore, that the percentage of teachers achieving the minimum acceptable level was reasonably high.

Many-Facet Rasch Analysis

With Many-Facet Rasch Analysis, it is possible to compute model-data fit of the different facets (assessors, teachers and assessment scales in this case) with rather less data than with traditional analyses such as correlations. See Appendix “Methodological Approaches and Analytical Tools” in Chap. 8 for a more detailed description and overview of Rasch measurement and Many-Facet Rasch Analysis.

Analysis of the three assessors indicated that all three emerged as lenient, although more importantly all showed acceptable model fit (Weigle, 1998).

A key issue concerns the extent to which the five levels which constitute each benchmark scale can be viewed as separate, or whether the levels overlap. If there is overlap, the boundaries between the levels will be perceived as tenuous. This is especially important at the crucial boundary between Levels 2 and 3—the benchmark level.

The analysis in Table 7.7 has also been computed using two standard errors, since this allows for a 95% level of confidence.

In order to reach the prototype benchmark level, teachers needed to score a ‘3’. When comparing the two crucial benchmark levels, it can be seen that Level 3 $-2SE$ had a measure of -0.30 logits, while Level 2 $+2E$ had a measure of -3.86 logits. This difference indicated that there was no overlap between Levels 2 and 3 on the scale. Model-data fit can be seen from the infit mean square figures in Column 8. Good fit—indicated by the infit mean squares in Column 8 being in the 0.5–1.5 range (Weigle, 1998)—could be noted at the crucial levels ‘2’ and ‘3’.

Table 7.7 Scale separation

1	2	3	4	5	6	7	8	← Column
Level	SE	-2SE	-1SE	Measure (logits)	+1SE	+2SE	Infit mean square	
5	0.11	7.53	7.64	7.75	7.86	7.97	1.7	
4	0.09	4.51	4.60	4.69	4.78	4.87	0.8	
3	0.15	-0.30	-0.15	0.00	0.15	0.30	0.7	Benchmark level
2	0.41	-5.50	-5.09	-4.68	-4.27	-3.86	1.6	
1				-7.75			3.0	

Further, it can be seen that at all levels, the +2SE figure of a given level is substantially below the -2SE figure of the level above, indicating that all the levels can be seen as distinct and separate.

Summary

The overall results of the PBAE for Classroom Language indicated that the majority of teachers of English in lower secondary classrooms would reach the minimum benchmark level if the ELBSC and ACTEQ were to recommend a Level '3' score on each component. It also indicated the following: that the assessors were reliable; that the scales and their descriptors worked well; that teachers' self-evaluations matched their assessment grades; and, unfortunately, that some teachers should perhaps not be teaching English.

The PBAE Speaking Test

Introduction

As reported in the validation study of the Speaking Test (Falvey & Coniam, 1999), the ELBSC agreed, after reviewing the different options proposed in the Consultants' Report for English Language Benchmarking (Coniam & Falvey, 1996), that the assessment of speaking was a crucial part of the English language benchmark assessment procedure. As stated earlier, the reason for retaining a speaking test was that teacher language performance was observed to be different, even on the same constructs, when language in the classroom is compared with the use of language with peers. In addition, the ELBSC eventually decided that some skills must be assessed for all teachers, e.g. the comparatively difficult and teacher-specific skill of reading aloud and the language teacher skill of storytelling or recounting. The Speaking Test that was agreed on for use in the PBAE consisted of three test types, assessed on six scales. Table 7.8 illustrates.

The descriptors used in the PBAE can be found Appendix E "Speaking Test Scales and Descriptors" in Chap. 5.

Implementation of the Speaking Test

The administration of the Speaking Test lasted four days, with two separate sessions on each day when different test material was used. There were twelve assessors in six teams. Each team consisted of two assessors.

Table 7.8 Scales and descriptors proposed by the ELBSC

Test type	Scale	Gloss
1. Reading aloud a text	1. Pronunciation, stress and intonation	Individual phonemes, stress, intonation
	2. Reading aloud with meaning	Speed of delivery, pausing, awareness of audience
2. Telling a story/recounting a personal experience/presenting arguments	1. Grammatical accuracy	Grammatical accuracy, range of structures
	2. Organisation and cohesion	Coherence, logical flow of ideas, relationships between ideas
3. Professional oral interaction	1. Interacting with peers	Including turn-taking, initiating, responding, agreeing and disagreeing
	2. Explaining language matters to peers	Including the use of appropriate metalanguage, appropriate examples

Analysis of Assessor Performance

Twelve assessors were used to assess test takers in the Speaking Test. Before they began to assess, they underwent a detailed training and standardisation programme, reported in Falvey and Coniam (1999). Using Multi-faceted Rasch analysis as the analytic statistic, all 12 assessors showed acceptable model fit, with a 1.4 logits leniency range +0.73 to -0.74.

Analysis of Test Taker Performance

The scores achieved by test takers are presented in different ways below. First, two sets of descriptive data are presented:

- the mean score for each scale
- a breakdown of the frequency scores for each scale (Table 7.9).

Compared with the CLA, where 84.4% of test takers achieved the benchmark grade of '3' on all four scales, in the Speaking Test, 41.6% achieved the benchmark Level of a '3' on all scales.

Table 7.9 Test takers achieving a 3 (or better)

On all 6 scales	126 (41.6%)
On 5 scales	39 (12.9%)
On 4 scales	42 (13.9%)
On fewer than 4 scales	95 (31.4%)

Table 7.10 Scales

Measure (logits)	Model error	Infit mean square	Assessors
+0.28	0.04	0.9	Pronunciation, stress and intonation
+0.19	0.04	0.9	Reading aloud with meaning
+0.12	0.04	0.9	Grammatical accuracy
-0.06	0.04	1.0	Organisation and cohesion
-0.19	0.04	1.2	Explaining language matters to peers
-0.35	0.04	1.1	Interacting with peers
0.00	0.04	1.0	Mean
+0.22	0.00	0.1	SD

RMSE 0.04; Adj S.D. 0.22; Separation 5.59; Reliability 0.97
 Fixed (all same) chi-square: 193.2; d.f.: 5; Significance: .00

Scale Difficulty and Separation

Table 7.10 presents an analysis of the six scales.

As can be seen from Table 7.10, all six scales had good infit mean square values [0.5–1.5 (Weigle, 1998)], with logit values clustered in a comparatively narrow range of just over half a logit.

Summary

It could be stated with reasonable confidence that for lower secondary (Years 7–9) teachers of English, the Speaking Test worked well. Some problems were identified and solutions sought; e.g., only one assessor (the interlocutor) should face the test taker; the other assessor should not be in the test taker's sight-line. And for Task 3 (group interaction) in order to facilitate interaction between the three test takers; the interlocutor too should move back to signal withdrawal from the three-way discussion.

The PBAE Writing Test

Introduction

A summary of the Writing Test is presented in Table 7.11.

Table 7.11 Scales and descriptors proposed by the ELBSC for the rewriting task

Test type	Scale	Salient linguistic features
1. Writing professionally	(1) Grammatical accuracy	Grammatical accuracy, range of structures
	(2) Organisation and coherence	Organisation of text, coherence
	(3) Task completion	All tasks requested in the stimulus must be completed
2. Rewriting a student composition	(1) Vocabulary and grammar	Grammatical accuracy, range of structures, appropriate lexical choice
	(2) Organisation and presentation of facts/information	Logical flow of ideas, relationships between ideas, retention of main facts/information from the original student text

Analysis of Assessor Performance

Three assessors were used to grade the Writing Test after a detailed training and standardisation programme (see Falvey & Coniam, 1999). After marking had been completed, any scripts that showed a discrepancy greater than two bands on the scale for any task, were given to a fourth assessor who re-graded the flagged scripts.

Many-Facet Rasch Analysis

An analysis of the three assessors indicated acceptable model-data fit. Measure values indicated that the three assessors were very similar in their awarding of grades; they all emerged as lenient, and in a narrow range—from -0.41 to -0.57 logits.

When comparing the two crucial benchmark levels, Level 3 $-2SE$ emerged with a measure of -0.59 logits, while Level 2 $+2E$ had a measure of -2.86 logits. This difference indicated that there was no overlap between Levels 2 and 3 on the scale. Separation could therefore be observed, indicating that the scales were sufficiently distinct for the effective operation of scales and the operationalisability of their descriptors.

Further, it can be seen that at all levels, the $+2SE$ figure of a given level is substantially below the $-2SE$ figure of the level above, indicating that all the levels could be perceived of as being separate, with no overlap between levels.

Analysis of Test Taker Performance

The scores achieved by test takers are presented in different ways below. It should be noted that on any scale, the band scores given by two assessors often results in a partial score, e.g. a score of ‘4’ from one assessor and a score of ‘3’ from the other assessor would result in a band score of 3.5 rather than a whole band score. Thus, in the tables below, the band scores are presented as score ranges.

In addition to the analysis presented as mean scores and percentages, some analysis is presented using multifaceted Rasch analysis, since this method of analysis attempts to put all elements in the assessment together on a common scale.

Table 7.12 presents the mean for each scale.

As can be seen from Table 7.12, the means for each scale are somewhat lower than the means awarded to the scales on Classroom Language Assessment. Many teachers appeared to have had too little time to complete the two tasks. The two tasks also indicate quite a spread of ability. Given that the ability of teachers ranged from educated native proficiency to those who are weak, a large SD on the Writing Test was, perhaps, to be expected.

While on Task 1 (Expository Writing) scores were clustered rather evenly at the ‘3’ and ‘4’ levels, on Task 2 (Rewriting), scores tended to be clustered rather more at the ‘2’ and ‘3’ levels, indicating that many more teachers appeared to be having problems with the Rewriting task.

Of all the test takers, only 7.6% (Task 1 (Expository Writing): *Organisation and coherence*) and 1.6% (Task 2 (Rewriting): *Grammar and vocabulary*) scored a ‘5’, indicating that teacher ability in writing was of a lower calibre than their ability in oral English (Table 7.13).

If teachers must pass all five scales of the Writing Test to pass the benchmark, 26.4% would have reached the benchmark. 19.1% achieved a ‘3’ on four of the five scales, and 20.4% obtained a ‘3’ on three scales. This suggests that either the written English of the majority of the cohort was below a minimum acceptable level, or that there were problems with test implementation, in particular, the amount of time allocated for this component. However, it should be noted that writing is the most difficult language skill at which ESL writers become proficient (see, for example, Bell & Barnaby, 1984; Bialystok, 1987; Nunan, 1989).

Table 7.12 Means for the five writing test scales

Scale	<i>N</i>	Mean	SD	Min.	Max.
Task completion	304	3.35	0.86	0	5
Organisation and coherence	304	3.39	0.89	0	5
Grammatical accuracy	304	2.83	0.92	0	5
Organisation and presentation	304	2.85	0.86	0	5
Vocabulary and grammar	304	2.62	0.81	0	5

Table 7.13 Test takers achieving a 3 (or better)

No. of scales	No. of test takers
On all 5 scales	79 (26.4%)
On 4 scales	57 (19.1%)
On 3 scales	61 (20.4%)
On 2 scales	48 (16.1%)
On less than 2 scales	55 (18.4%)

Summary for the Writing Test

A number of issues arose from the implementation of the PBAE Writing Test. Test takers found that the Writing Test was the most difficult of the criterion-referenced instruments. It was not clear whether this was because of the problems second language speakers (and, incidentally, native speakers too) have with writing, the unfamiliarity of test takers with Task 2 and its implied dual purpose, or the relatively short amount of time allocated to the test.

The PBAE Reading Test

This section describes the results obtained in the PBAE Reading Test. In the Reading Test, the 19 questions which constituted the test have been broken down into 39 smaller items because some questions required a number of points to be included in the answer. The results are presented in Table 7.14.

The mean for the Reading Test was 0.46 (SD=6.66)—slightly more difficult than the multiple-choice cloze test, whose mean was 0.59. The Reading Test mean was also very close to that obtained in the Reading Test validation study (Coniam & Falvey, 1998a, b, c, d) where the mean was 0.41. This would suggest that the difficulty level of the PBAE Reading Test was appropriate for a test intended for use with teachers.

The majority of the items worked well. Item 22 required test takers to deduce attitude. It also required test takers to include two pieces of evidence to support their answer. One of the Reading Test markers commented here that:

Many candidates cannot deduce the attitude of the Principal from what he says and the word, 'disgruntled'.

Table 7.14 Test results for the reading test

Number of items	39
Mean	17.9 (45.9%)
Standard deviation	6.66 (17.1%)
Alpha	0.83
Standard error of measurement	2.75 (7.1%)
Mean point biserial correlation	0.48
Good item (good discrimination, good facility)	27
Acceptable item, although easy (good discrimination, high facility)	
Acceptable item, although difficult (good discrimination, low facility)	5
Acceptable item (acceptable discrimination, good facility)	4
Marginally acceptable item (acceptable discrimination, low facility)	2
Poor item (low discrimination, low facility)	1

Very few can include both pieces of criticism in their answer, indicating a lack of thoroughness.

The worst item was item 12 (0.11 correct; PBC 0.18), which was too difficult and did not discriminate. As can be seen, however, a test with only one poor item out of 39 is an indication that the test has worked well.

According to Ebel (1965, p. 337), the expected reliability for a 40-item test should be in the region of 0.67. This is clearly achieved by the PBAE Reading Test with a reliability coefficient of 0.83. The standard error of measurement was 2.75, or 7.1%. According to Donlon (1984) as a general guide, the SEM should be below 10%. The standard errors of measurement of the SAT verbal and quantitative scores have been reported to be in the range of 29–34 points (Donlon, 1984, pp. 33–34), or 4.4–5.7% of the score range.

While only a small number of test takers scored the maximum possible on questions which were marked 2-1-0 or 3-2-1-0, there was a good spread of scores across the possible range of scores within a question, indicating that test takers were able to answer part, if not all, of most questions. One-third of the questions had a 10% omission rate—suggesting that there were perhaps too many questions, or that the questions were too demanding. This issue was one that would need to be addressed in future administrations of the benchmark test.

Table 7.15 Test results for the cloze test

Number of items	21
Mean	12.5 (59.5%)
Standard deviation	3.31 (15.8%)
Alpha	0.68
Standard error of measurement	1.87 (8.9%)
Mean point biserial correlation	0.51
Good discrimination, good facility	12
Acceptable discrimination, high facility	6
Acceptable discrimination, low facility	3
Unacceptable items	–

The Multiple-Choice Cloze Test

The Multiple-choice Cloze Test was included in the battery of tests in order to provide an ‘anchor’ against which other test instruments might be compared. The results for the Cloze Test are provided in Table 7.15.

The mean for the multiple-choice Cloze Test was 0.60 ($SD = 3.31$). This compares very favourably with the pretested cloze tests (Coniam & Falvey, 1998a) where the means were 0.64 and 0.60, respectively. This would suggest that the difficulty level is appropriate for a test destined for use with teachers.

As can be seen from Table 7.15, the quality of the items produced was generally high. In the current cloze test, it was decided that all the items had worked well enough not to warrant deletion. Item performance was better than in the validation study of the multiple-choice cloze (Coniam & Falvey, 1998a), where some items required deletion from the cloze passages.

The PBAE Listening Test

Introduction

The format for the Listening Test developed by the ELBSC involved answers in an open-ended format, which can be described as a hybrid between a test for English language teachers, and, as mentioned, a more demanding HKASLE Year 13 Use of English Listening Test. This is because the majority of the questions generally paralleled the text but some did not, with the result that test takers had to draw on what they had heard from different sections of the taped discussion. For purposes of marking, and in a manner similar to the grading of the Reading Test, it was agreed that marking should reflect an understanding of content and its inferences only.

The Evaluation Study

The Listening Test consisted of an audio-taped discussion between two male speakers, on which 19 open-ended questions had been set.

Test Statistics

The test was treated as comprising 70 items, so that a classical item analysis could be conducted. The results are presented in Table 7.16.

As can be seen from Table 7.16, 24 of the 70 items could be classed as good, while another 30 were acceptable. Of the remainder, 10 items were poor. Another six were marginal; however, their facility values were so low, that they would have to be deleted if the test were a multiple-choice test.

According to Ebel (1965, p. 337), the expected reliability for an 80-item test is 0.80. This is clearly achieved by the PBAE Listening Test with a reliability coefficient of 0.83 for 70 items. The standard error of measurement was 3.30, or 4.7%, well within acceptable limits (see Donlon, 1984, pp. 33–34 for a discussion of the SEM in the SAT verbal and quantitative scores).

However, in spite of the high standards of reliability, the acceptable SEM described above and the consequent 54 good items, it was apparent that the Listening Test emerged as too difficult—with a mean of 0.33.

Table 7.16 Test results for the listening test

Number of items	70
Mean	22.9 (32.7%)
Standard deviation	9.99 (14.2%)
Alpha	0.89
Standard error of measurement	3.30 (4.7%)
Mean point biserial correlation	0.48
Good item (good discrimination, good facility)	24
Acceptable item, although easy (good discrimination, high facility)	1
Acceptable item, although difficult (good discrimination, low facility)	23
Acceptable item (acceptable discrimination, good facility)	6
Marginally acceptable item (acceptable discrimination, low facility)	6
Poor item (low discrimination, low facility)	10

Table 7.17 Correlations between test types

	Reading	Writing	Speaking	Listening	CLA	
Cloze	0.541	0.596	0.576	0.590	0.340	PPM
	0.000	0.000	0.000	0.000	0.000	Sig. (2-tl'd)
Reading		0.544	0.522	0.669	0.385	PPM
		0.000	0.000	0.000	0.000	Sig. (2-tl'd)
Writing			0.696	0.655	0.543	PPM
			0.000	0.000	0.000	Sig. (2-tl'd)
Speaking				0.595	0.613	PPM
				0.000	0.000	Sig. (2-tl'd)
Listening					0.483	PPM
					0.000	Sig. (2-tl'd)

PPM Pearson correlation

Correlations with Other Test Types

Table 7.17 presents the correlations between test types in the PBAE battery of tests.

Correlations were generally high between all test types, apart from the CLA, where a moderate (although still significant) correlation existed. Listening and Reading correlated quite highly at 0.67, and Speaking and Writing at 0.70.

Test Taker Reactions to the Different Test Components

After finishing each PBAE test component, test takers were asked to complete a questionnaire in order to obtain feedback.

In general, three questions on the questionnaire asked test takers how they felt about the different test components. Generally, these questions concerned:

1. How valid they felt the component tasks were
2. How easy or difficult they found the component tasks
3. How they rated their performance on the component tasks.

Reactions to the Classroom Language Assessment Test

The questionnaire was completed by 278 of the test takers. After completing the questionnaire, qualitative data was sought from approximately 10% of the cohort (30 teachers) who gave in-depth interviews about their reactions to benchmarking, the PBAE tests and its administrative procedures.

Table 7.18 Digest of written comments

Category	Type of comment	Number
Own performance	Made mistakes in grammar	2
	Nervous being observed	6
Expectations	Unsure what assessor is looking for	3
Feedback	Not provided after assessment—would be welcome	5
Use of Chinese	Should I use it to help understanding—especially in low-band schools	2
Students	Were nervous	2
	Low standard (band 5) who are less responsive	5
Assessors	One observation not enough to say if can teach suitably	4

On the issue of the authenticity of the CLA, respondents reacted favourably in their perception, with 45.6% responding that they felt the Classroom Language Assessment was an authentic means of assessment. Written comments which test takers chose to add were on the whole either critical or defensive. Comments in Table 7.18 are reported on the basis of similar comments being made by more than one respondent.

Reactions to the Speaking Test

On the issue of the validity of the Speaking Test, respondents were divided in their reaction to the test's validity, with 24.3% responding that they felt the Speaking Test was valid; 42.7% had no opinion, and 31.6% felt that the Speaking Test task was invalid. It is unlikely that the majority of respondents had the metalanguage to understand what 'validity' means in terms of assessment. On the question of personal performance on the Speaking Test, respondents were also balanced in their answers. 19.4% felt that they had performed below average, 59.0% felt they had performed averagely, and 20.7% felt the test had been easy. Written comments which test takers chose to add were, in the main, critical (Table 7.19).

Table 7.19 Digest of written comments on the reading test

Category	Type of comment	No. of comments
Time	Not enough time to finish	6
Subject matter	Good passage/relates to career	2
	Part 2 very demanding and difficult	4
	Ambiguous questions	4
	Most Chinese don't use that kind of language daily	3
	Requires specific knowledge in that area	2

Reactions to the Writing Test

On the issue of the validity of the Writing Test tasks, respondents were quite categorical in their views of the validity of the test. The majority agreed that the test was valid.

On the question of test difficulty, opinions illustrated the same weighting. 29.0% of respondents felt that the test was easy; 56.0% had no opinion, and 12.7% felt it was difficult.

On the question of personal performance on the Writing Test, 14.3% of respondents felt that they had performed below average, 59.3% felt they had performed in-between, and 24.0% felt the test had been easy. Written comments which test takers chose to add were mixed. Seventeen test takers commented that there had not been enough time to finish the Writing Test section. While seven test takers commented that the Writing Test was 'very practical and authentic', five felt that it was 'unrealistic to re-write students' work'. Two test takers commented on the inappropriacy of the Writing Test for Primary teachers of English.

Reactions to the Reading Test

On the issue of the validity of the Reading Test tasks, respondents reacted favourably, with 35.3% responding positively; 42.0% had no opinion, and 20.7% felt that the reading task was invalid.

On the question of test difficulty, opinions were again very balanced. 22.6% of respondents felt that the test was easy; 56.7% had no opinion, and 18.7% felt it was difficult.

On the question of personal performance on the Reading Test, respondents were also balanced in their answers. Written comments which test takers chose to add were, in the main, critical.

Reactions to the Listening Test

In contrast with the Reading Test, where half of the responses fell in the middle 'in-between' score, few respondents hedged. Most responded either positively or negatively on the issue of test validity, generally negatively.

On the first question above, respondents were split fairly evenly. On the second question of level of difficulty, opinions were more consistent than on any other question: 86.3% responded that the test was difficult.

Responses to the question of personal performance on the Listening Test mirrored test takers' perception of the level of difficulty of the test, with 76% feeling that they had performed below average.

A large number of respondents commented on: poor sound quality; problems with the format; difficulty in making out what speakers said; and difficulties in differentiating between the speakers. This suggests that, given the low mean, the difficulty that test takers experienced with the Listening Test was due to the nature and quality of the Listening Test that was produced for the PBAE, not necessarily to low ability in listening as a skill.

Summary

This section has focused on the validation process of the different components' of the PBAE through the analysis of test takers' and assessors' performance, Many-Facet Rasch Analysis, an analysis of correlation of different test components and an analysis of test takers' reaction to different components. The analysis shows that the Level 3 was a suitable benchmark level for Hong Kong English language teachers. Chapter 8 describes the sensitive and difficult issue of determining benchmarks, once the pilot study had been completed and analysed.

References

- Bell, J., & Barnaby, B. (1984). *A handbook for ESL literacy*. Toronto: OISE.
- Bialystok, E. (1987). A theoretical language learning model. *Language & Learning*, 28, 69–83.
- Coniam, D., & Falvey, P. (1996). *Setting language benchmarks for English language teachers in Hong Kong secondary schools*. Hong Kong: Advisory Committee on Teacher Education and Qualifications.
- Coniam, D., & Falvey, P. (1998a). *Validating the classroom language assessment component: The Hong Kong English language benchmarking initiative*. Hong Kong: Advisory Committee on Teacher Education and Qualifications.
- Coniam, D., & Falvey, P. (1998b). *Validating the speaking test: The Hong Kong English language benchmarking initiative*. Hong Kong: Advisory Committee on Teacher Education and Qualifications.
- Coniam, D., & Falvey, P. (1998c). *Piloting the multiple-choice cloze test: The Hong Kong English language benchmarking initiative*. Hong Kong: Advisory Committee on Teacher Education and Qualifications.
- Coniam, D., & Falvey, P. (1998d). *Validating the writing test: The Hong Kong English language benchmarking initiative*. Hong Kong: Advisory Committee on Teacher Education and Qualifications.
- Donlon, T. (1984). *The college board technical handbook for the scholastic aptitude test*. New York: College Entrance Examination Board.
- Ebel, R. L. (1965). *Measuring educational achievement*. Englewood Cliffs, NJ: Prentice-Hall.
- Falvey, P., & Coniam, D. (1999). *Assessor training and standardisation for classroom language assessment: The Hong Kong English language benchmarking initiative*. Hong Kong: Advisory Committee on Teacher Education and Qualifications.
- Hatch, E., & Lazaraton, A. (1991). *The research manual*. Boston, MA: Heinle and Heinle.
- Nunan, D. (1989). *Designing tasks for the communicative classroom*. Cambridge: Cambridge University Press.

- Rosenthal, R., & Rosnow, R. L. (1975). *The volunteer subject*. New York: Wiley.
- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 263–287.

David Coniam is Chair Professor and Head of Department of the Department of Curriculum and Instruction in the Faculty of Education and Human Development at The Education University of Hong Kong, where he is a teacher educator, working with teachers in Hong Kong primary and secondary schools. His main publication and research interests are in language assessment, language teaching methodology and computer assisted language learning.

Peter Falvey is a teacher educator, formerly a Head of Department in the Faculty of Education, The University of Hong Kong. His main publication and research interests are in language assessment, first and second language writing methodology, and text linguistics.

Chapter 8

Determining Benchmarks After the PBAE



David Coniam and Peter Falvey

Abstract This chapter discusses how benchmark levels were determined after the results of the PBAE were made available to the ELBSC (English Language Benchmark Subject Committee); it examines how analytically marked tests may be calibrated with criterion-referenced tests and how ‘cut’ scores may be determined for them.

Determining Benchmark Levels and ‘Cut’ Scores

One of the major issues examined in this chapter is the need to determine a ‘passing score’ or ‘cut score’ as the benchmark for each of the assessment instruments which have been developed. A benchmark score may be determined in advance for each of the criterion-referenced tasks in the battery. Then, when there is a mixture of criterion-referenced tasks and analytically marked tests, a common scale can be derived for the analytically marked tests by aligning them with one of the criterion-referenced instruments, thus identifying ‘cut scores’. The ‘passing score’ which then corresponds to the benchmark need not be the same for each of the assessment instruments involved (e.g. the mean). With regard to the criterion-referenced tests, the ELBSC recommended—through an analysis of desired performance—that the benchmark passing score be set at Level ‘3’ on each of the five-point scales. The analytically marked tests—the Reading and Listening Tests—however, proved to be a problem. One of these various methods of test equating uses *expert judgement* (see Nedelsky 1954; Angoff 1984). Essentially, these methodologies involve experts reviewing test content in order to calculate the degree of agreement arrived at on item and test difficulty. In reviewing Angoff’s (1984) approach of using expert judgement,

D. Coniam (✉) · P. Falvey
Department of Curriculum and Instruction, Faculty of Education and Human Development,
The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, Hong Kong
e-mail: coniam@eduhk.hk

P. Falvey
e-mail: falvey@eduhk.hk

Fulcher and Davidson (2007) point out that such an approach is relatively easy to carry out and helps with defining a 'minimally competent master', although they acknowledge that such an approach may encounter validity problems when judges differ widely in their decisions. In the field of English language, Stansfield, Karl, and Kenyon (1990) used the expert judgement method on the setting of four tests of reading and cloze for *The Guam Educators' Test of English Proficiency*. As stated earlier, this method is described in much greater detail by Drave in Section III of this volume.

Whereas Fulcher and Davidson (2007) consider using expert judgement is more suitable for dichotomously scored items (i.e. items which are scored as either right or wrong), the possibility of using expert judgement in open-ended questions has been explored. Buck (1991) was among the early researchers who investigated the use of expert judgement in short-answer comprehension questions on an English language listening comprehension test. In an analysis of the results of 20 experts (applied linguists and assessment personnel), there was strong agreement among the experts on the ratings given to items tapping lower-level processing, but less agreement on items tapping higher-level processing. In a study examining students' different responses to different text types, Kobayashi (2002) found that a majority of the expert assessors in the study were able to identify the text types of more than two-thirds of the passages.

Another procedure involves the use of Rasch measurement (IRT values) to align the scores of the analytically marked tests with those of the criterion-referenced tests. In this procedure, the reference scale would be one representing a teacher's ability to teach in English. Two procedures were used in the PBAE to investigate how passing scores might be derived for the analytically marked tests—expert judgement and aligning using Rasch measurement.

Modelling Cut Scores

Alignment by Means of Expert Judgement

An adaptation of the Angoff method (1984), which uses expert judgement, was adopted to determine the benchmarks of the three papers which did not use grade descriptors, namely the Reading Test, the Listening Test and the multiple-choice Cloze Test.

The first set of tables provided by the HKEA below present the means for the three tests.

The expert judges' perceived means for what would constitute a passing score in Table 8.1a were quite dissimilar, ranging from 0.28 to 0.49. This contrasted with the item mean for the Listening Test of 0.14. The discrepancy between the two sets of scores would suggest that on the basis of the judges' perceptions virtually no one

Table 8.1 a Listening Test: expert judges’ mean scores, b Reading Test: expert judges’ mean scores, c Cloze Test: expert judges’ mean scores

	N	Min.	Max.	Mean	SD
<i>a) Listening Test</i>					
A	47	0.00	1.00	0.47	0.28
B	47	0.10	0.90	0.46	0.26
C	47	0.00	1.00	0.28	0.41
D	47	0.10	0.90	0.48	0.24
E	47	0.10	0.80	0.49	0.18
F	47	0.10	0.80	0.48	0.22
G	47	0.00	0.80	0.36	0.21
Item facility	52	0.00	0.54	0.14	0.12
<i>b) Reading Test</i>					
A	36	0.10	0.80	0.46	0.18
B	36	0.10	0.80	0.38	0.19
C	36	0.00	1.00	0.47	0.31
D	36	0.00	0.80	0.34	0.18
E	36	0.10	0.90	0.41	0.18
F	36	0.10	0.90	0.47	0.21
G	36	0.10	0.90	0.47	0.20
Item facility	37	0.02	0.78	0.31	0.17
<i>c) Cloze Test</i>					
A	34	0.20	0.80	0.65	0.15
B	34	0.30	0.90	0.59	0.14
C	34	0.00	1.00	0.68	0.28
D	34	0.20	0.90	0.64	0.21
E	34	0.50	0.90	0.75	0.11
F	34	0.40	0.90	0.68	0.16
G	34	0.60	0.90	0.77	0.10
Item facility	34	0.16	0.92	0.63	0.21

should have passed the Listening Test, given that the actual mean, as mentioned, was 0.14.

The results of expert judges in the Reading Test—where the range of the judges’ scores was 0.34–0.47—were similar to the Listening Test although less harsh. The Reading Test mean was again lower than that of the most lenient judge—judge D, who estimated that a mean of 0.34 would be necessary to be benchmarked on the Reading Test.

The Cloze Test emerges as the test type where the judges’ opinions most closely matched item difficulty—a range of 0.59–0.77 against an actual test mean of 0.63.

Item-Person Interaction				
Items	Persons			
	Location	Std Error	Location	Std Error
Mean	0.000	0.189	-0.239	-0.040
SD	1.447	1.826	0.860	0.881
Correlations	0.118	0.086		
Complete data degrees of freedom = 284.82			171.09	
Item-Trait Interaction				
Total Item Chi Sq	1373.283	Person separation index	0.963	
Total Degree Freedom	688.000	Cronbach	N/A	
Total ChiSq Probability	0.000			
Test of Fit Power		EXCELLENT		

Fig. 8.1 First reports the overall test calibration results

Summary: Expert Judges' Ratings

The results of the experts' judgements were quite varied. This phenomenon matches the criticisms of some researchers in Cizek's overview (1996a) of the expert judgement approach to standard setting, particularly those of Shepard (1984). See also Drave in Section III of this volume for a further discussion of this issue.

Alignment Using Rasch Measurement

This section describes the use of Rasch measurement to align radically different types of approaches to assessment. As mentioned earlier, in the PBAE, the latent criterion was taken as a teacher's language ability to teach in English, with the different assessment instruments therefore viewed as different manifestations of that ability. The procedure can therefore be seen to be a valid means of attempting to set benchmarks for other test instruments in the PBAE.

In order to test the requirement for unidimensionality, test calibration involved attempting to align the six test types (Classroom Language Assessment, Speaking, Writing, Reading, Listening and Cloze) onto a single scale. Test calibration was then performed as follows via the Rasch Unidimensional Measurement Model (RUMM) package (Andrich, Lyne, Sheridan, & Luo, 1997):

1. Overall conformity of the test items to the Rasch model was first estimated.
2. Degrees of conformity to the Rasch model for each test item were derived.
3. The six subtests were finally aligned by identifying the test scores which corresponded to the same ability.

The following tables and figures present the results of calibration.

Figure 8.1 first reports the overall test calibration results.

The overall fit of the calibration, as indicated in Fig. 8.1, was *Excellent*, according to RUMM.

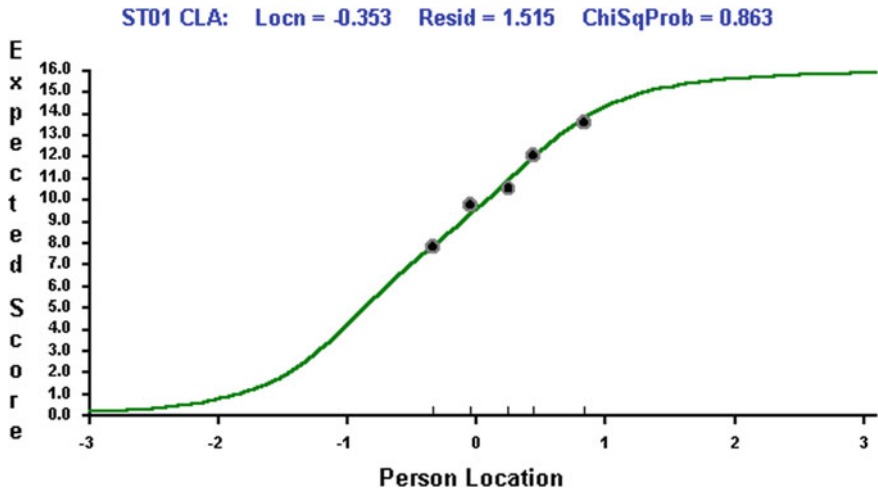


Fig. 8.2 Calibration for Classroom Language Assessment

For all the subtests, model fit was high, indicated by the fact that, for all tests, on each figure the groups (the dots in the figures) appeared almost exactly along the line of the curve, corroborating the summary analysis in Table 8.1 of ‘excellent’ power of fit.

A sample is presented for the calibration for Classroom Language Assessment only, to give the reader a visual presentation of the analysis (Fig. 8.2).

Aligning Tests and Identifying Benchmark Levels

Apart from calibrating the results of tests, *RUMM* also made it possible to align test score points onto the Rasch scale. Table 8.2 reports part of the results of the test equating. The table is laid out so that the score down the first column is the score on the Classroom Language Assessment component. The arrow against ‘12’ represents the prototype benchmark level for Classroom Language Assessment (assuming that level ‘3’ was required for each of the four scales).

It can be seen from Table 8.2 that the score from the Rasch calibration analysis which corresponded to ‘12’ in Classroom Language Assessment was 0.29 logits. ‘12’ in this instance refers to the minimum standard (Level 3) being scored on each of the four scales—hence 4×3 making 12. This score of 0.29 logits (or the nearest logit value) could be taken as the potential benchmark score for the other test types. In order not to fail teachers who would otherwise be benchmarked, it was proposed that it would be advisable to take the score at the lower end of a particular ability band.

Table 8.2 Test aligning results—midpoint ability scores

Score	CLA	Reading	Listening	Cloze
Percent equivalent—scale 'mid-point'	80.00%	57.89%	42.03%	70.00%
1	-4.77	-4.02	-4.19	-4.6
2	-4.01	-3.28	-3.46	-3.77
3	-3.5	-2.83	-3.02	-3.24
4	-3.07	-2.49	-2.69	-2.81
5	-2.67	-2.22	-2.43	-2.45
6	-2.27	-1.99	-2.21	-2.12
7	-1.86	-1.79	-2.02	-1.81
8	-1.42	-1.6	-1.85	-1.52
9	-0.99	-1.44	-1.69	-1.25
10	-0.56	-1.28	-1.55	-0.97
11	-0.14	-1.13	-1.42	-0.7
12 →	0.29	-0.99	-1.29	-0.43
13	0.74	-0.85	-1.17	-0.16
14	1.29	-0.72	-1.06	0.12
15	2.11	-0.6	-0.95	0.42
16		-0.47	-0.84	0.75
17		-0.35	-0.74	1.11
18		-0.23	-0.65	1.53
19		-0.11	-0.55	2.07
20		0.02	-0.46	2.89
21		0.14	-0.37	
22		0.26	-0.28	
23		0.38	-0.19	
24		0.51	-0.11	
25		0.64	-0.02	
26		0.77	0.06	
27		0.91	0.14	
28		1.05	0.22	
29		1.2	0.3	
30		1.36	0.38	
31		1.54	0.46	
32		1.72	0.54	

(continued)

Table 8.2 (continued)

Score	CLA	Reading	Listening	Cloze
33		1.93	0.62	
34		2.17	0.7	
35		2.44	0.78	
36		2.79	0.86	
37		3.24	0.94	
38		3.99	1.02	
39			1.1	
40			1.18	
41			1.26	
42			1.35	
43			1.43	
44			1.52	
45			1.6	
46			1.69	
47			1.78	

Determining an Overall Benchmark Level for the PBAE Tests

Introduction

In their deliberations, both ACTEQ and the ELBSC made constant references to the ways in which a benchmark ‘pass’ might be arrived at, viz.:

Whether a teacher *must* ‘pass’ *all* components of *every* benchmark (i.e. CLA, Speaking Test, Reading Test [both reading and cloze], Listening Tests and Writing Test) in order to receive an overall benchmark grading.

Whether a teacher *must* ‘pass’—within each test type—*all* the relevant scales of a certain test component. Consider the Speaking Test, for example, which consists of six scales. In order to be benchmarked, i.e. achieve a ‘pass’, on the Speaking Test, test takers would have to achieve the benchmark level of 3 on all six scales.

This section therefore reports how many teachers would pass on each separate test and on the overall battery of assessment instruments.

After much deliberation, the ELBSC recommended that passes—for the PBAE—would be determined as follows:

- (1) the criterion-referenced scale-based tests, i.e. CLA, the Speaking, the Writing Test.

On these double-marked tests, the ELBSC recommended that a pass should be determined as reaching the benchmark level (i.e. Level ‘3’) on each scale except for one, where an average level of 2.5 would be permitted.

Table 8.3 Test takers passing the different test components

Test component	Pass mark	Pass mark (%)	Max poss.	Passing	Percent (%)
CLA	11.5	47.9	24	259/302	85.8
Speaking	17.5	58.3	30	183/303	60.4
Writing	14.5	58.0	25	119/300	39.7
Listening	13	18.6	70	260/297	87.5
Reading	11	28.2	39	265/298	88.9
Cloze	9	42.9	21	272/299	91.0

Table 8.4 Test takers reaching the 'overall' benchmark for English

No. of tests on which the benchmark was reached	Number	Percent
1. CLA	10	3.4
2. Speaking	16	5.4
3. Writing	38	12.9
4. Listening	58	19.7
5. Reading	78	26.5
6. Cloze	93	31.6
Total	294	100.0

(2) the analytically marked tests, i.e. the Listening, Reading and Cloze Tests

On these tests, the ELBSC accepted that Rasch measurement should be used to determine the cut score, using test takers' scores on the CLA as the anchor.

The ELBSC also recommended that in order to be 'benchmarked', test takers would need to pass *every* test.

Calculation of Benchmarks

Table 8.3 summarises the information presented so far in this chapter, laying out what percent of teachers would be deemed to have passed each test component of the PBAE.

Table 8.4 now presents the results for the overall benchmark. The figures have been calculated for those test takers who took all six components of the PBAE.

As can be seen from Table 8.4, if test takers needed to pass *every* section of *every* component test in order to be 'benchmarked', only 93 out of the total cohort of 294, or 31.6%, would reach the prototype benchmark. The following table presents the picture of test takers' results by qualification on the different tests.

As Table 8.5 illustrates, all groups performed comparatively well on the CLA and the analytically marked tests. The best performing group nonetheless was Group

Table 8.5 Test takers by qualification for each test

	Primary or secondary; academic and/or professional qualifications	CLA	ST	WT	LT	RT	CT
1	Pri—T cert, subject-trained	25/29 (86.2%)	11/29 (37.9%)	2/28 (7.1%)	21/28 (86.2%)	21/28 (86.2%)	21/28 (86.2%)
2	Pri—T cert, non-subject-trained	19/25 (76.0%)	6/25 (24.0%)	1/25 (3.8%)	13/23 (86.2%)	15/24 (86.2%)	20/24 (86.2%)
3	Sec—T cert, subject-trained	46/51 (90.1%)	32/51 (62.7%)	19/51 (37.2%)	45/51 (86.2%)	44/51 (86.2%)	44/51 (86.2%)
4	Sec—T cert, non-Subject-trained	9/12 (75.0%)	6/12 (50.0%)	1/12 (8.3%)	11/12 (86.2%)	10/12 (86.2%)	10/12 (86.2%)
5	Sec—degree, relevant	27/30 (90.0%)	24/30 (80.0%)	17/30 (56.7%)	28/30 (93.3%)	29/30 (96.7%)	29/30 (96.7%)
6	Sec—degree, relevant, PG prof trg	30/31 (96.8%)	30/33 (90.1%)	25/33 (75.8%)	31/33 (93.9%)	32/33 (97.0%)	32/33 (97.0%)
7	Sec—degree, non-relevant	61/78 (78.2%)	39/77 (50.6%)	27/75 (36.0%)	67/74 (90.5%)	69/74 (93.2%)	70/75 (93.3%)
8	Sec—degree, non-relevant, PG prof trg	42/46 (91.3%)	35/46 (76.0%)	27/46 (58.7%)	44/46 (95.6%)	45/46 (97.8%)	46/46 (97.8%)
		259/302 (85.8%)	183/303 (60.4%)	119/300 (39.7%)	260/297 (87.5%)	265/298 (88.9%)	272/299 (91.0%)

pri primary; *sec* secondary; *t cert* teachers’ certificate; *PG prof trg* postgraduate professional training

6—professionally qualified secondary school teachers with a relevant degree. The two worst-performing groups were Groups 2 and 4—teachers without a relevant English language qualification who were not professionally qualified.

Tables 8.6 and 8.7 further highlight the differences between Group 6 and the other groups.

Table 8.6 illustrates that the group which achieved the highest number of ‘passes’ were secondary school teachers who held a relevant degree and were professionally qualified (Group 6). In this group, 29 out of 33 (87.9%) achieved the benchmark specified as ‘all elements must be passed to achieve the benchmark’. The second highest scoring group, (Group 8), was secondary school teachers holding a non-relevant degree, but still professionally qualified. In this group, 38 out of 46 (82.6%) would have reached the benchmark. Of the primary school teachers in the cohort PBAE, 8 out of 51 (15.7%) would have reached the benchmark.

Table 8.7 is a summary of test takers who passed in all six test components and those who passed in five out of six components.

Table 8.7 illustrates that the group which achieved the highest number/percentage of ‘passes’ was secondary school teachers who held a relevant degree and was professionally qualified (Group 6). In this group, 21 out of 33 (67.7%) achieved the benchmark specified as ‘all elements must be passed to achieve the benchmark’.

Table 8.6 Test takers reaching benchmark

Grp	Pri/sec	Acad/prof qualification	N	No. of tests in which benchmarked					
				1	2	3	4	5	6
1	Pri	T cert, subject-trained	28	1	4	4	12	5	1
2	Pri	T cert, non-subject-trained	23	5	3	5	4	5	1
3	Sec	T cert, subject-trained	51	2	2	7	9	19	12
4	Sec	T cert, non-subject-trained	12		2	2	4	3	1
5	Sec	Degree, relevant	30			4	2	10	14
6	Sec	Degree, relevant, PG prof trg	33			2	2	6	21
7	Sec	Degree, non-relevant	73	2	5	9	20	18	19
8	Sec	Degree, non-relevant, PG prof trg	46			5	5	12	24
			294	10	16	38	58	78	93

pri primary; *sec* secondary; *t cert* teachers' certificate; *PG prof trg* postgraduate professional training

Table 8.7 Test takers reaching the benchmark in all tests

Grp	Pri/sec	Acad/rof qualification	Passing in five test components (%)	Passing in all six test components (%)
1	Pri	T cert, subject-trained	21.4	3.6
2	Pri	T cert, non-subject-trained	26.0	4.3
3	Sec	T cert, subject-trained	60.8	23.5
4	Sec	T cert, non-Subject-trained	33.0	8.3
5	Sec	Degree, relevant	80.0	46.7
6	Sec	Degree, relevant, PG prof trg	87.1	67.7
7	Sec	Degree, non-relevant	50.7	26.0
8	Sec	Degree, non-relevant, PG prof trg	78.3	52.2

Legend pri = primary; sec = secondary; t cert = teachers' certificate; PG prof trg = postgraduate professional training

Summary of Test Taker Performance

On the individual test types, the range of test takers who reached the prototype benchmark ranged from 39.7% on the Writing Test to 91.0% on the Cloze Test.

On the basis of being benchmarked (meaning that all tests in the battery must be passed), it has been illustrated that 31.6% of test takers on the PBAE would be declared 'benchmarked'. A final discussion of how well a benchmark pass related to test takers' academic and professional background and qualifications revealed that

the most able group of teachers in the PBAE was secondary school teachers who held a relevant degree and were professionally qualified.

The PBAE: Internal and External Perspectives of the Reliability and Validity of Its Results, and the Representative Nature of the Sample

Before discussing conclusions and recommendations, this section examines the constitution of the sample of teachers that participated in the piloting of the assessment instruments for the English language benchmark initiative in Hong Kong. A discussion is then presented of the extent to which the sample may be seen to have been representative of Hong Kong teachers of English. Two issues are addressed in this section:

- The extent to which the results of the complete PBAE cohort might be viewed as reliable and valid for all participants in the PBAE—an *internal* perspective.
- The extent to which the results of the PBAE may be taken as representative of the wider context of English language teachers in Hong Kong—an *external* perspective.

The Representative Nature of the PBAE Sample

A major issue regarding the representative nature of the PBAE sample [discussed in depth in Coniam and Falvey (2002)] concerned willingness to participate and volunteering. Coniam and Falvey (*ibid*) report that of the PBAE sample, approximately 50% reported that they had volunteered of their own free will to participate, while 50% stated they had been ‘persuaded’ or ‘instructed’ to participate. Test takers’ reactions to the test types indicated that willing volunteers were, on the whole, more positively inclined towards the benchmark test types than were unwilling test takers who had been instructed to participate. In a comparison of the test scores of willing volunteers with those of unwilling non-volunteers, however, little in the way of significance emerged between the test scores for these two different groups. This lack of significance suggested therefore that the PBAE results—for which the sample of teachers included those with a wide range of ability and qualifications—could therefore be taken as being representative of Hong Kong secondary English language teachers.

The Representative Nature of the PBAE Vis-a-Vis the General Hong Kong English Language Teacher Cohort

The study by Coniam and Falvey (2003) attempted to establish a point of reference between test takers who sat the PBAE tests and the broader population of Hong Kong lower secondary English language teachers. The comparison was made possible because a number of robust test items which had been calibrated with a representative sample of more than 10,000 Hong Kong Years 7–13 secondary school students (Coniam, 1995) had been administered to PBAE test takers along with the five subtests which constituted the benchmark test. Not only had the items been calibrated against a large number of Hong Kong English language students, a subset of the items had been included in certain local universities' entrance tests to Postgraduate Certificate in Education (PGCE) programmes in English language teaching, the latter making it possible to extend the Years 7–13 scores further by two more grades—'notional' Years 14 and 15, which would then begin to encompass teacher ability score levels.

An examination of samples in both the 1998 PBAE and a 1998 PGCE entry test showed that both groups scored comparably—a t test run between both groups' results showed no significant difference. While the calibrated items were discrete-point multiple-choice lexico-syntactic/usage items, they could only be taken as a snapshot of the similarities between the two groups' abilities in that they did not reflect the range of abilities which the different benchmark subtests were designed to sample. Nonetheless, the fact that the two groups' results were not significantly different lent support to the assertion that the PBAE sample was reasonably indicative of the ability of the general population of Hong Kong lower secondary English language teachers. The mean for the whole cohort of PBAE test-takers (excluding Groups 4 and 6 who were demonstrably better than the rest of the PBAE cohort) was slightly above that of a notional Year 15 score (notionally equivalent to the second year of a university undergraduate programme) and similar to that of PGCE applicants. The external reference-point of the calibrated items thus performed (to a limited extent) the role of a standardised test.

The PBAE—Conclusions and Recommendations

This section summarises the conclusions and recommendations arising from the administration of the PBAE. They relate essentially to the test battery and the general implementation and administration of the benchmark tests.

Recommendations with Regard to Specific Tests

Speaking Test

Despite the reservations of some teachers in the pilot exercise about reading a poem in addition to a prose passage, the ELBSC felt that the poem should be retained, mainly because of the positive washback effects it would have in the language classroom. In the setting of assessment tasks, while setters and moderators should be reasonably sensitive to the possible ramifications of certain issues, (e.g. divorce), no topics should be automatically avoided as long as they are broadly related to education and professional language teaching. A more detailed discussion of the use of a poem in the speaking test and its dropping in the revision of the LPATE in 2007 is provided in the closing Chap. 18.

Writing Test

While Task 2 (rewriting a student composition) might be further improved, the ELBSC recommended that the task should be retained, but be further split into two separate tasks consisting of:

- (i) identification of problems
- (ii) a rewrite to produce a model essay

It was also suggested that the time for the Writing Test (of both Tasks 1 and 2) be increased from 75 to 90 min.

Listening Test

Although the ELBSC felt that eventually, when the technology had matured, video would be a better means of input for the Listening Test, they agreed that audio input should be continued for the next few years. (For a discussion of the use of video or audio as input for the Listening Test, see Coniam, 2001 and Chap. 18).

Reading Test

The ELBSC noted the problems of grading answers to open-ended questions and felt that the test might be assessing more than comprehension because the grammaticality of answers might influence the marker. Despite this, it was agreed that open-ended questions should be retained.

Table 8.8 Cut scores for the criterion-referenced tests

	CLA	Writing	Speaking
No. of scales	4	5	6
Benchmark level	11.5	14.5	17.5

Cloze Test

The analysis of the multiple-choice test in the PBAE revealed that most test takers who did well in the Reading Test did similarly well in the MC Cloze Test. The ELBSC recommended that a pass in both the Reading Test and MC Cloze Test would be required to meet the benchmark.

General Test Recommendations

Test Specifications of the Assessment

The ELBSC recommended that all the criterion-referenced tests and all the paper-and-pencil tests should be retained, at least for the first live implementation of benchmark assessment. The method of using calibrated multiple-choice items should be kept as ‘anchors’.

Calculating a Benchmark ‘Pass’ on the Criterion-Referenced Tests

In determining the ‘pass rates’ in the criterion-referenced tests, the following recommendations were made by the ELBSC:

Where two assessors are used, the mean scores of the two assessors should be taken as the test taker’s score for each scale.

The cut score for each criterion-referenced test would be determined as achieving the benchmark level, i.e. a ‘3’ on each scale, with ‘2.5’ permitted on one scale. Effectively, the ‘cut scores’ are the minimums laid out in Table 8.8.

Aligning Paper-and-Pencil Tests with Criterion-Referenced Tests

In principle, the Rasch measurement model should be used. It was also decided that analysis using expert judgement would be undertaken to satisfy those who strongly supported the ‘expert judgement’ method.

Pretesting

All tests should be pretested before being administered.

General Recommendations

Exemptions

At this stage of the process, no exemptions were recommended for any group of teachers, because of the possible difficulties in defining what the ‘relevant degrees’ were.

Briefing and Preparation for Teachers

To help teachers better prepare for the assessment, it was recommended that exemplar materials (e.g. audio/video recording of benchmarkable teachers being assessed for CLA, or tapescripts of lessons, etc.) should be made available. Another point that must be clearly communicated to teachers was the issue of teachers overdoing the use of Cantonese in class, so that there was not enough evidence to make a judgement on the teacher’s language ability.

Responsibility for Benchmarking Pre-service Teachers

The ELBSC’s view was that even for the pre-service teachers, all the paper-and-pencil tests, including the Speaking Test, should be conducted in a centralised manner by the HKEA (or similar body). In addition, the ELBSC suggested that, resources permitting, the assessment of pre-service teachers for CLA should also be conducted by the same team of EDB assessors assessing the in-service teachers.

Test Administration

It was proposed that all teachers would have to take the same set of tests, except for Speaking, which would be conducted over a few weeks if necessary.

It was felt that Saturdays were in general preferable to holidays, but since all the paper-and-pencil tests would take a day and a half, or three mornings in total, administering the tests only on Saturday mornings would effectively deprive teachers of three weekends.

Access to Teachers' Profile/Results

This was an issue only for the live assessment. The teachers taking the live assessment should be given a profile of their performance and not just a result slip saying whether they had been benchmarked or not. This profile should be given to teachers as soon as possible, subject to logistic arrangements, e.g. depending on the size of the candidature—so that they receive useful feedback on their performance.

Summary

The current chapter concludes Part I and describes how LPATE cut scores were decided through a modelling process involving both expert judgement and Rasch measurement. The chapter concluded with recommendations regarding the different test components and general test administration.

Section I now ends with Appendix “[Methodological Approaches and Analytical Tools](#)”, entitled *Methodological Approaches and Analytical Tools*. This describes the major approaches and tools used in Part I in order to better inform readers.

Section II follows with a description of the courses that were created to assist teachers to develop their skills and reach adequate benchmark levels.

Appendix: Methodological Approaches and Analytical Tools

This Appendix may be viewed as a helpful aid to readers between the background and the context of the benchmark initiative described in this book. For the initiated, the section may appear somewhat lightweight, while for those less well versed in assessment principles and techniques, it might appear to be too difficult so an attempt has been made to provide something accessible to everyone. While knowledgeable readers will want to skip elements of this section, further references have been provided for relative newcomers. The section attempts to provide an overview of the methodological approaches, both qualitative and quantitative, employed in the benchmark/LPATE research studies. It further describes the analytical tools used to augment those methodologies. Readers are introduced to quantitative survey approaches and the classical statistics used in those approaches. Studies reported draw on both Classical Test Theory as well as Rasch measurement, the latter enabling different facets (e.g. person ability and item difficulty) to be modelled together. Rasch analysis helps to provide better assessments of performance, enhances the quality of measurement instruments and provides a clearer understanding of the nature of the latent trait (Bos, Goy, Howie, Kupari, & Wendt, 2011). Better-informed readers may wish to refer to sources such as Bachman (2005) or Green (2013), which more fully explain some of the statistical issues detailed in Appendix.

Methodological Approaches

1. Quantitative survey approach
2. Qualitative analysis including a Grounded Theory approach

Quantitative Survey Approach

A quantitative survey approach employs questionnaires as the main source of data collection. It is important to note that for purposes of reliability and validity, questionnaires should go through an iterative process in order to ensure that the questionnaires that eventually are given to prospective respondents are the best that can be made available (e.g. Dornyei, 2003). This is done through a process of drafting, scrutiny, piloting, analysing, re-drafting, re-piloting and eventually administering before final analysis (see Brown, 2001, pp. 7–12).

Qualitative Analysis Including a Grounded Theory Approach

A qualitative approach to data collection makes use of qualitative data collected during written responses to questions and oral responses in one-to-one interviews, group interviews and the oral data from focus groups.

A *Grounded Theory* approach (Glaser & Strauss, 1967) is commonly used with qualitative data analysis. It consists of various iterations of analysis involving qualitative data. Examples of qualitative data are the product of open interviews (e.g. one-to-one interviews), what is said in group interviews or focus groups and the written responses provided by respondents to questionnaires which allow free responses to written stimuli. Grounded Theory allows researchers to discover theory from data which has been obtained systematically and then analysed to look for patterns.

Methodological Tools

This section describes the use made of Classical Test Statistics, Rasch measurement, Rasch models and qualitative data analysis.

Certain studies described in this book have, in the main, used Classical Test Theory (CTT) to analyse data—specifically survey data. While the use of CTT enables statistical significance to be examined, there are inherent weaknesses with CTT statistics. First, analytical techniques in CTT require linear, interval scale data input (Wright, 1997). Raw data collected through Likert-type scales, however, are usually ordinal since the categories of Likert-type scales indicate only ordering without

any proportional levels of meaning. Applying conventional analysis on ordinal raw data can therefore lead to potentially misleading results (Bond & Fox, 2007; Wright, 1997). Second, CTT uses total score to indicate respondent ability levels. This results in person ability estimates being item-dependent; i.e., although person abilities may be the same, person ability estimates are high when items are easy but low when items are difficult. Similarly, item difficulty estimates are similarly sample-dependent; i.e., even though item difficulties themselves are invariant, item difficulty estimates appear high when respondents' competence is low but low when respondents' competence is high.

Classical Test Theory (CTT)—often called the 'true score model'—assumes that every test taker has a true score on an item if it is possible to measure that score directly without error. CTT analyses assume, therefore, that a test taker's test score is comprised of a test taker's 'true' score plus a degree of measurement error.

An overview of the CTT statistics used in the current set of studies will be briefly presented below. These can be grouped broadly into **Descriptive Statistics** (statistics that simply describe the group that a set of persons or objects belong to) and **Inferential Statistics** (statistics that may be used to draw conclusions about a group of persons or objects).

Descriptive statistics used in the studies are the **mean** (the arithmetical average), the **standard deviation** (the measure of variability in the dataset) and the **variance** (the average of the squared differences from the mean; the standard deviation squared, in effect.).

Inferential tests may be conceived of as either **parametric** or **nonparametric**. **Parametric data** has an underlying normal distribution—which allows for greater conclusions to be drawn since the shape can be described in a more mathematical manner. Other types of data are all **nonparametric**.

Parametric and Nonparametric Tests

Parametric Tests

Parametric inferential statistical tests used in the case study have been the t test, ANOVA and Pearson correlations. These will now be briefly described.

The T Test

The t test is used to compare two population means, with a view to determining if there is a significant difference between the means. There are two types of t tests, **unpaired t tests** (where the samples are independent of one another) and **paired t tests** (where the samples are related to each other). A t test is commonly used when

the variances of two normal distributions are unknown and when an experiment uses a small sample size [a sample size of 30 subjects is used in the studies as being the threshold for conducting statistical analysis (Ramsey, 1980)].

Analysis of Variance (ANOVA)

ANOVA is used to compare differences of means among more than two groups. This is achieved by looking at variation in the data and computing where in the data that variation occurs (giving rise to the name 'ANOVA'). Specifically, ANOVA compares the amount of variation between groups against the amount of variation within groups.

The Pearson Product-Moment Correlation (PPM)

The Pearson correlation is an estimate of the degree of the relationship between two variables. The scale runs from -1 through 0 to $+1$, where $+1$ shows a total positive correlation, 0 indicates no correlation, and -1 shows a total negative correlation.

The inter-rater correlation is one application of the PPM, indicating the measure of agreement between raters of scale-based assessment. Interpretations of correlation magnitude differ. Friedrich (1999), for example, suggests that a correlation of 0.5 indicates a 'moderate to strong tendency'. Hatch and Lazaraton (1991, p. 441) suggest a 'strong' correlation, as regards inter-rater reliability, be taken as 0.8 . Following the example of Friedrich (1999) and Hatch and Lazaraton (1991), a correlation of 0.5 has been adopted in these studies to indicate a moderate correlation, one between 0.5 and 0.8 as moderate to strong, and a correlation above 0.8 as strong.

Nonparametric Tests

The nonparametric inferential statistical test used in the case study has been the chi-squared test.

The Chi-Squared Test

The chi-squared test is used with *nominal* data (where the data fall into 'categories', for example male/female, or Likert scales in the current studies). The chi-squared tests compare the counts of responses between two or more independent groups and

determine whether there is a significant difference between expected and observed frequencies in one or more category.

Significance

All the statistical tests described above—both parametric and nonparametric—provide a figure regarding the level of significance (the p value) which emerged on the test. The p value is the probability of the result occurring by chance or by random error. The lower the p value, the lower is the probability that the event being measured can be explained by chance. A p value lower than 5% ($p < 0.05$) is generally accepted as the threshold of statistical significance, although in many cases the 1% level ($p < 0.01$) indicates a stronger case for arguing for significance (see Whitehead, 1986, p. 59). A p value > 0.05 therefore suggests no significant difference between the means of the populations in the sample, indicating that the experimental hypothesis should be rejected. Over the past few decades, there have been a number of controversies about the use/over-use of significance in data analysis. A useful overview is provided in Glaser (1999, pp. 291–296) and Schneider (<https://arxiv.org/ftp/arxiv/papers/1402/1402.1089.pdf>—accessed July 2017).

Test and Test Item Statistics

Facility Index

The range for an item with acceptable facility is taken as being in the range of 0.3–0.8. (see Falvey, Holbrook, & Coniam, 1994, p. 119ff)

Discrimination Index

An item discrimination (the point biserial correlation) of above 0.3 is considered ‘good’. A discrimination of 0.2–0.3 is considered ‘workable’ while a discrimination of below 0.2 is considered unacceptable. (See Falvey, Holbrook, & Coniam, 1994, p. 126ff)

Test Reliability

Cronbach's alpha is a test reliability statistic which is generally the starting point for determining a test's worth, with the desirable level (for longer tests, i.e. 80 or more items) usually taken as 0.8 (see Ebel, 1965, p. 337). With shorter tests, lower reliability figures are cited; Ebel (1965, p. 337), for example, states 0.6 for 30 items.

Test Mean

An ideal mean for a 'final achievement' test (Hughes, 2003, p. 13) should be in the region of 0.5. Such a mean suggests—as Gronlund (1985) comments—that the test is generally appropriate to the level of a 'typical' or 'average' student in the class or group. A low mean can suggest that the test is too difficult, with a high mean suggesting that it is too easy (Zimmerman, Sudweeks, Shelley, & Wood, 1990, p. 10). A mean in the region of 0.5 in general indicates that most students managed to finish it, i.e. that they did their best and did not simply guess. Further, a mean of 0.5–0.6 indicates that student scores are spread out and maximises a test's discriminating power (Gronlund, 1985, p. 103).

Standard Error of Measurement

The standard error of measurement (SEM) indicates the extent to which test scores match 'true' scores because all tests will contain a degree of error. As a general rule, an SEM below 10% might be considered desirable. On the controversial Massachusetts Teacher Tests quite a large SEM (17%) was reported—see Haney, Fowler, Wheelock, Bebell, and Malec (1999) for a discussion of the problems associated with the administration of the Massachusetts Teacher Tests—which may be why opponents of the test felt that its reliability was questionable.

Effect Size

While statistical differences are discussed in terms of statistical significance, standard deviation units (SDUs) are also provided in certain instances so that the size of the differences between the two groups may be appreciated. Following Cohen (1988, pp. 477–478), an SDU of 0.2 indicates a small effect, 0.5 a medium effect and 0.8 a large effect.

The Rasch Model and Many-Facet Rasch Analysis

It should be noted that the following sections on the Rasch Model and Many-Facet Rasch Analysis are not dissimilar to other descriptions of the Rasch Model and Many-Facet Rasch Analysis in some of Coniam and Falvey's previously published articles (see, for example, Coniam & Falvey, 2016, 52–55; Coniam & Falvey, 2001). This is because such descriptions do not need to vary much and often—as below—use the same metaphor of Rasch measurement being rather like using a ruler in order to place results on it for measurement purposes. They may appear to be rather long for the knowledgeable reader, but an attempt has been made to cover the two topics thoroughly.

In contrast to CTT, the use of the Rasch model (1960, 1980) enables different facets (e.g. person ability and item difficulty) to be modelled together. First, in the standard Rasch model, the aim is to obtain a unified and interval metric for measurement. The Rasch model converts ordinal raw data into interval measures which have a constant interval meaning and provide objective and linear measurement from ordered category responses (Linacre, 2006). This is not unlike measuring length using a ruler, with the units of measurement in Rasch analysis (referred to as 'log-its') evenly spaced along the ruler. Second, once a common metric is established for measuring different phenomena (test takers and test items being the most obvious), person ability estimates are independent from the items used, with item difficulty estimates being independent from the sample recruited because the estimates are calibrated against a common metric rather than against a single test situation (for person ability estimates) or a particular sample of test takers (for item difficulty estimates). Third, Rasch analysis prevails over CTT by calibrating persons and items onto a single unidimensional latent trait scale—also known as the one-parameter IRT (Item Response Theory) model, (Bond & Fox, 2007; Wright, 1992). Latent Trait Analysis (LTA), a form of latent structure analysis (Lazarsfeld & Henry, 1968), is used for the analysis of categorical data. Person measures and item difficulties are placed on an ordered trait continuum by which direct comparisons between person measures and item difficulties can be easily conducted. Consequently, results can be interpreted with a more general meaning. Further, as the Rasch model provides a great deal of information about each item in a scale, its use enables the researcher to better evaluate individual items and how these items function in a scale (Törmäkangas, 2011).

The Rasch model has been widely applied in educational research, especially in the field of large-scale assessment (Schulz & Fraillon, 2011; Wendt, Bos, & Goy, 2011). It helps to provide better assessments of performance, enhances the quality of measurement instruments and provides a clearer understanding of the nature of the latent trait (Bos, Goy, Howie, Kupari, & Wendt, 2011).

Many-Facet Rasch Analysis (MFRA) and Data Analysis

MFRA refers to a class of measurement models that extend the basic Rasch model by incorporating more variables (or facets) than the two that are typically included in a test (i.e. test takers and items), such as markers, scoring criteria and tasks.

In Hong Kong English language public examinations, test takers' final grades are computed directly from markers' raw scores. While the latter may be adjusted for mean and standard deviation on the basis of correlations with other papers taken by the test takers, essentially the result is the raw score. The accuracy of the information obtained from raw scores has long been questioned, with the problems associated with their use discussed by a number of researchers, with a number of studies commenting on how the use of raw scores constitutes an imperfect measure of test taker ability (McNamara, 1996, p. 122; Weir, 2005). Weir (2005), discussing scoring validity with the need for test results to be as free as possible from measurement error, stable and consistent over time and reliable, states '... if FACETS [a Many-Facet Rasch Analysis computer program] is not being used in the evaluation of writing tests, I would want to know why not!' A study which examined the use of raw scores in the application of rating scales in the HKCEE 2005 Writing Test (Coniam, 2008) illustrated how the use of raw scores and measures derived through MFRA could produce markedly different results for test takers.

As described earlier, in the area of language performance tests (see, e.g., McNamara, 1996, p. 9), the major statistical method of analysis accepted over the past decade has come to be MFRA, since it allows for situational factors such as marker severity, prompt difficulty to be modelled and compensated for (McNamara, 1996, p. 4; Weir, 2005, p. 199). It should be noted that McNamara considers productive English language speaking and writing tests as weak versions of such tests.

Overall data-model fit in MFRA can be assessed by examining the responses that are unexpected given the assumptions of the model. According to Linacre (2006), satisfactory model fit is indicated when about 5% or less of (absolute) standardised residuals are equal or greater than 2 and about 1% or less of (absolute) standardised residuals are equal to or greater than 3.

One of the key statistics in MFRA is the infit mean square statistic. This describes model fit, with 'fit' essentially being the difference between expected and observed scores. Definitions of fit vary. 'Perfect fit' according to Bond and Fox (2007, pp. 285–286) is defined as 1.0, with an acceptable upper limit of fit stated as 1.3. Weigle (1998) proposes acceptable practical limits of fit as 0.5 for the lower limit and 1.5 for the upper limit.

Constructs, Scales, Descriptors and Benchmarks

Unlike objective tests or tests which are analytically marked, and for which measures of reliability and discrimination can be found using the types of statistical packages

described above, criterion-referenced assessment instruments follow different patterns of quality assurance, procedures, validity and reliability.

The process begins when a *construct* (the underlying skill, knowledge or performance which is to be measured or assessed) is identified. Bachman and Palmer (1996) state that the identification of a construct is reached through a process of description of the purpose of the test; a description of the *target language use* (TLU); and a description of the test takers. This activity allows for the description of 'the precise nature of the ability we want to measure' which can be defined abstractly, thus creating a theoretical definition of the construct which 'provides the basis for considering and investigating the construct validity of the interpretations we make of test scores' (ibid, pp. 88–89).

Once the *construct* has been identified, the creation of *scales* for each construct is attempted. *Scales* are the component of assessment design which allows raters to make decisions on the test takers performance. They can be created either dichotomously, with only two *levels*—e.g. 'can achieve/cannot achieve the benchmark' or, alternatively, by creating different *levels* for each *scale*, which is a more complex undertaking. North (2000) states that:

Scales of language proficiency have become relatively widespread over the past decade as part of a general movement towards more transparency in educational systems, which places a higher value on being able to state what the attainment at a given level of language proficiency means in practice (2000, p. 9)

Each *level* of a scale indicates the quality of test taker response to the task or activity they have been asked to attempt. In order to judge the quality of a test taker's response, each level requires a description against which the rater can match the test taker's performance. The descriptions may relate to the skills, knowledge or performances that are underpinned by the theoretical construct and its scales. These descriptions, which are normally text-based, are called *descriptors*. They are intended to provide all stakeholders with a clear, accessible understanding of what is required by the test takers.

Benchmarks are arrived at ideally when all those who are stakeholders in the assessment development process agree or determine the level on each scale which is accepted by all as the minimum skill/piece of knowledge/performance which must be achieved by test takers. Benchmarks can be adjusted upwards on a scale or downwards on the scale depending on what policy-makers decide is the appropriate minimum level.

Rating scales which have been developed for purposes of performance assessment may be classified in different ways. These involve holistic or analytic, primary or multiple trait perspectives (Hamp-Lyons, 1991; Weigle, 1998), or from what might be referred to as real-world or ability/interaction perspectives (Bachman, 1990). In terms of use, the orientation of rating scales may then be either towards users, towards assessors or towards the test constructor themselves (Alderson, 1991). Broadly two major approaches to rating scale design are identified in the literature.

The first, and most established, approach comes from a theoretical perspective and draws on the constructs or abilities to be measured. Under this approach, the rating

scale is designed on the basis of a measurement model determined by experts in the field. Experts consider samples of performance data as a post hoc activity—selecting and identifying samples that typify specific levels on a scale or domain.

The second approach is grounded in a more empirical perspective—making use of written or oral texts produced by learners which are taken as exemplars of performance. Turner and Upshur's (2002) binary-choice, boundary definition scales are possibly one of the most cogent examples of this empirical approach. Under this method, the scale and the ensuing cognitive process that raters must adhere to are laid out as a set of repeated branching binary decisions.

Fulcher, Davidson, and Kemp (2011) criticise measurement-driven scales—the first approach described above—as suffering from ‘descriptive inadequacy’, arguing that scales derived in this manner are insufficiently sensitive to communicative context or the complexities of interaction that are typically associated with language use. They suggest that levels of abstraction in such scales are broad and create gulfs between a score and what the score represents. Making a case for richer descriptions of contextually based performance, which they argue strengthen the meaning of a score, and the validity of inferences which may consequently be derived from such scores, they propose the use of performance decision trees—an extension of Turner and Upshur's (2002) boundary definition scales.

Background to Standards Setting

The sections below describe standard setting, especially within shifting paradigms and discuss the notion of validity in standards setting. Further discussion of this topic in relation to the LPATE is provided by Drave in Chap. 14.

Changing Attitudes to Standards Setting Within Shifting Paradigms

Cizek's (1996a) comprehensive survey of the literature on setting standards and the methods employed to set standards is still valid even though it was conducted some 20 years ago. It is summarised below for reference.

Early work on standard setting was based on the principle that there was a ‘right answer’ based on population parameters and that ‘it is the task of standard setting to find it’ (Jaegar, 1989, p. 492). The certainty about standard setting had shifted by the late 1970s where some viewed standard setting as both arbitrary and capricious (Glass, 1978, pp. 253, 258). Cizek, however, cites Block (1978) and Popham (1978) as among those who opposed this view. They felt that ‘standard setting was not an arbitrary process, or, at least, that it was not arbitrary in the sense of being capricious’. An emerging view rejected the population parameter method with, by 1984, Shepard

stating that ‘The standard we are groping to express is a psychological construct in the judges’ minds’ (1984, p. 188).

In the 1990s, two definitions of standard setting emerged to replace the ‘parameter estimation perspective’. They were, first, Cizek’s (1993) procedural definition of standard setting that ‘focuses on the process’—defined as ‘the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance’. Here, the focus is on ‘a process that can be used to rationally derive, consistently apply and explicitly describe procedures by which inherently judgmental decisions are made’ (Cizek, 1996b, p. 21). Second was Kane’s definition that proposed that the process of standard setting ‘... draw(s) a distinction between the passing score, defined as a point on the score scale, and the performance standard, defined as the minimally adequate level of performance for some purpose ... The performance standard is the conceptual version of the desired level of competence, and the passing score is the operational version’ (1994, p. 426).

Cizek makes it clear that most of the procedures and rules mentioned are based on discrete-item tests and that the standard setting mechanisms devised by the American Educational Research Association (AERA), American Psychological Association (APA) and the National Council on Measurement in Education (NCME) are designed to prevent bias and to create as much validity as possible for these indirect tests of ability. However, some of the five basic guidelines have considerable merit and include, as their primary (basic) standards:

- Standard 1.24—describes how rates of misclassification will vary depending on the percentage of individuals tested who belong to each category.
- Standard 5.11, 8.6, 10.9—makes information available regarding the rationale of the test and a summary of the evidence supporting intended interpretations, including evidence about the validity of the cut.
- Standard 6.9—provides details on the standard-setting method used and the rationale for setting a cut score, including information about the qualifications of the participants in the process.

Cut scores are selected points on the score scale of a test. The points are used to determine whether a particular test score is sufficient for some purpose. For example, student performance on a test may be classified into one of several categories such as basic, proficient, or advanced on the basis of cut scores (Zieky & Perie, 2006).

As can be seen, most of these guidelines apply to discrete-item tests where a raw or standardised score results from sitting the test but where the ‘value’ of that score is not known except in terms of whether the candidate had passed/failed the test. Three of the English language benchmark tests have scales and descriptors (the validation and assessor training of which are described in subsequent chapters of this report) which have been arrived at by due process, are already transparent to test-takers and the public and contain the ‘value’ of the score in the descriptor that accompanies each scale. It was with this in mind that the English Language Consultancy Team and the English Language Benchmark Subject Committee (ELBSC) at that time turned to the Code of Practice of ALTE (The Association of Language Testers in Europe) as

best representing the work that the benchmark project was attempting to carry out. ALTE was chosen because, besides focusing on language assessment, it has also always worked with a mixture of psychometric and performance testing (witness its long tradition of oral examining and language teacher assessment).

In this context, it is important to note that authorities such as UCLES (now known as Cambridge Assessment) have, over decades, consistently used direct tests of language ability (where, for example, strong emphasis is placed on the training and standardisation of assessors to enhance reliability in their oral assessments). Such tests include CEELT (the Cambridge examination in English for Language Teachers—see Falvey & Andrews, 1994) and Cambridge's range of language proficiency tests such as the First Certificate in English (FCE), the Preliminary English Test (PET) and the hundred-year-old Cambridge Proficiency in English Test (CPE).

The American-led tradition, however, has until relatively recently focused on issues of reliability in their indirect measurement tests (driven by issues of fairness and possible litigation) while ignoring and excluding major issues of validity raised through concerns engendered by the testing of performance by indirect methods (Cizek, 1993, 1996a, b; Greenburg, 1992). As their tests have tended to be paper-and-pencil norm-referenced tests, they have had to focus on validity constructs (required because it is notoriously difficult to prove the validity of indirect tests of psychological constructs) rather than validity itself which comes from measurement of performance in a direct test of ability.

Four of ALTE's major guidelines for the setting of standards for language are:

1. Provide prompt and easily understood reports of examination results that describe candidate performance clearly and accurately.
2. Describe the procedures used to establish pass marks and/or grades.
3. If no pass mark is set, then provide information that will help users follow reasonable procedures for setting pass marks when it is appropriate to do so.
4. Warn users to avoid specific, reasonably anticipated misuses of examination results.

The ALTE guidelines (see <http://www.alte.org>) are not, in essence, different from those of the AERA, APA and NCME. They insist on proper procedures and the provision of accounts of the 'due process' that was entailed in setting up the constructs, setting specifications, setting and moderating tests, piloting them, sampling test-takers and explaining to test-takers and other stakeholders what these processes are and how the passing grades have been arrived at.

Much of this 'due process' was detailed in the first consultancy report (Coniam & Falvey, 1996), eight subsequent reports and their account of the formal pilot, the Pilot Benchmark Assessment (English) (PBAE).

It should be noted that over the past twenty years, Cizek has added to his 1996 comments by showing, in 2007 (Cizek and Bunch), that essentially the same methods that he cited in 1996 still apply (e.g. the application of Angoff; Nedelsky and Ebel) plus others such as *The Direct Consensus Method* (see Pitoniak, 2003); *The Contrasting Groups and Borderline Group Methods* (see Humphrey-Murto, & MacFadyen, 2002); *The Body of Work and Other Holistic Methods* (Wyse, Bunch,

Deville, & Viger, 2014); Hambleton and Pitoniak (2006). The Item-Descriptor Matching Method (see Ferrara et al., 2008) and The Hofstee and Beuk Methods (see Wyse & Babcock, 2017).

By 2000, Hambleton (2000) was stating that methods for setting performance standards on educational assessments using the multiple-choice item format were well developed and steps for implementation were generally clear (see Livingston & Zieky, 1982). On the other hand, he stated that standard-setting methods for educational assessments that include constructed response items such as writing samples and performance tasks were not as well developed at that time, and none of them had been fully researched. He added that new methods, improved implementation of existing methods, and increased efforts to validate any performance standards were needed.

The Notion of Validity in Standards Setting

On the issue of validity in standard setting, Cizek stated (1996b) that validity “*does not exist outside of the values systems that define what are desirable outcomes: What is considered ‘reasonable’ or ‘appropriate’ ultimately depends on individual values*” (p. 28).

Messick (1989) attempted to bring some order to this process by proposing a framework for validating standards where an ongoing process of gathering and evaluating evidence helped standard setters to focus on whether the inferences implied by application of a cutting score are warranted. The ‘due process’ of reaching consensus on the criterion-referenced tests by involving large numbers of potential stakeholders took place throughout the English language benchmark process, not only with the PBAE but with studies before the PBAE and since the PBAE, all of which helped to throw light on the benchmarks that were being developed and the standards that were being ‘recommended’.

Examples of those involved included ACTEQ members, consisting of both professional and lay members, PGCE English majors, Refresher Training Course members at The University of Hong Kong, Masters students (English Majors) at The University of Hong Kong, staff at City University, Hong Kong Polytechnic University, The University of Hong Kong, The Chinese University of Hong Kong and the Hong Kong Institute of Education and thousands of teachers who attended seminars organised by EMB.

In the Consultancy Team, the principal investigators were from The Chinese University of Hong Kong and The University of Hong Kong with other consultants/investigators drawn from other secondary, vocational and tertiary institutions, as well as from the UK.

The consultancy team was constituted in order to reflect a broad spectrum of expertise and experience in a number of relevant areas:

- Local and international language teacher education experience and expertise
- Local and international language teaching experience and expertise at primary and secondary levels
- Local and international language educational assessment experience and expertise
- Local and international language educational assessment administrative and organising experience and expertise
- Local primary and secondary English teaching.

In the context of setting standards and recommendations, it should be noted that test developers and subject committees only ‘recommend’ standards. The final decision is always taken by another stakeholder body—which in this instance was ACTE-Q—which takes into account all other factors which a test developer might ignore (such as financial and human resources, socio-political factors, teacher readiness).

References

- Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s*. London: Modern English Publications and the British Council.
- Andrich, D., Lyne, A., Sheridan, B., & Luo, G. (1997). *Rasch unidimensional measurement models (RUMM) computer program*. RUMM Laboratory: Murdoch University, Australia.
- Angoff, W. H. (1984). *Scale, norms, and equivalent scores*. Princeton, N.J.: Educational Testing Service.
- Bachman, L. (2005). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Block, J. H. (1978). Standards and criteria: A response. *Journal of Educational Measurement*, 15, 291–295.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: fundamental measurement in the human sciences* (2nd ed.). Mahwah, N.J.: Erlbaum.
- Bos, W., Goy, M., Howie, S. J., Kupari, P., & Wendt, H. (2011). Rasch measurement in educational contexts Special issue 2: Applications of Rasch measurement in large-scale assessments. *Educational Research and Evaluation*, 17(6), 413–417.
- Brown, J. D. (2001). *Using surveys in language programs*. New York, NY: Cambridge University Press.
- Buck, G. (1991). The testing of listening comprehension: An introspective study. *Language Testing*, 8(1), 67–91.
- Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement*, 30, 93–106.
- Cizek, G. J. (1996a). Standard-setting guidelines. *Educational Measurement: Issues and Practice*, 15, 13–21.
- Cizek, G. J. (1996b). Standard-setting guidelines. *Educational Measurement: Issues and Practice*, 15, 13–21.

- Cizek, G. J. (1996c). An NCME instructional module on: Setting passing scores. *Educational Measurement: Issues and Practice*, 15(2), 20–31.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications Ltd.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Coniam, D. (1995). Towards a common ability scale for Hong Kong English secondary school forms. *Language Testing*, 12(2), 184–195.
- Coniam, D. (2008). An investigation into the effect of raw scores in determining grades in a public examination of writing. *Japan Association for Language Teaching Journal*, 30(1), 69–84.
- Coniam, D. (2001). The use of audio or video comprehension as an assessment instrument in the certification of English language teachers: A case study. *System*, 29(2), 1–14.
- Coniam, D., & Falvey, P. (1996). *Setting language benchmarks for English language teachers in Hong Kong secondary schools*. Hong Kong: Advisory Committee on Teacher Education and Qualifications.
- Coniam, D., & Falvey, P. (2001). Assessor training in a high-stakes test of speaking: The Hong Kong English language benchmarking initiative. *Melbourne Papers in Language Testing*, 8(2), 1–19.
- Coniam, D., & Falvey, P. (2002). The representative nature of a sample: The Hong Kong Pilot Benchmark Assessment (English) exercise. *Hong Kong Journal of Applied Linguistics*, 7(1), 16–33.
- Coniam, D., & Falvey, P. (2003). Benchmarking the benchmark: Assessing the fit of a new test with its target population of teachers of English in Hong Kong. *Hong Kong Journal of Applied Linguistics*, 8(1), 1–15.
- Coniam, D., & Falvey, P. (2016). The Hong Kong education and examination system. In D Coniam (Ed.), *English language education and assessment: Recent developments in Hong Kong and the Chinese Mainland* (pp. vi–vii). Singapore: Springer.
- Dornyei, Z. (2003). *Questionnaires in second language research: Construction, administration, and processing*. Mahwah, NJ: Lawrence Erlbaum.
- Ebel, R. L. (1965). *Measuring educational achievement*. Englewood Cliffs, NJ: Prentice-Hall.
- Falvey, P., & Andrews, S. (1994). The Cambridge examination in English for language teachers (CELT). In J. Boyle & P. Falvey (Eds.), *English language testing in Hong Kong*. Hong Kong: The Chinese University Press.
- Ferrara, S., Perie, M., & Johnson, E. (2008) Matching the judgmental task with standard setting panelist expertise: The item-descriptor (id) matching method. *Journal of Applied Testing Technology*, 9(1), 1–22.
- Falvey, P., Holbrook, J., & Coniam, D. (1994). *Assessing students*. Hong Kong: Longman.
- Friedrich, K. (1999). *Interpreting correlation coefficients*. <http://acad.cl.uh.edu/itc/educ6032/courses/resources/unit2/index.htm>. October 8 1999.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. London & New York: Routledge.
- Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, 28(1), 5–29.
- Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine.
- Glaser, D. N. (1999). The controversy of significance testing: Misconceptions and alternatives. *American Journal of Critical Care*, 5(5), 291–296.
- Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement*, 15(4), 237–261.
- Green, R. (2013). *Statistical analyses for language testers*. Basingstoke: Palgrave Macmillan.
- Greenburg, K. (1992). Validity and reliability issues in the direct assessment of writing. *WPA Writing Program Administration*, 16(1–2), 7–22.
- Gronlund, N. E. (1985). *Measurement and evaluation in teaching*. New York: Macmillan.

- Hambleton, R. K. (2000). Setting performance standards on educational assessments and criteria for evaluating the process. In G Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Publishers.
- Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: Praeger Publishers.
- Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241–276). Norwood, NJ: Ablex.
- Haney, W., Fowler, C., Wheelock, A., Bebell, D., & Malec, N. (1999). Less truth than error? An independent study of the massachusetts teacher tests. *Education Policy Analysis Archives*, 7(4). <http://epaa.asu.edu/epaa/v7n4/>.
- Hatch, E., & Lazaraton, A. (1991). *The research manual*. Boston, MA: Heinle and Heinle.
- Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.
- Humphrey-Murto, S., & MacFadyen, J. C. (2002). Standard setting: A comparison of case-author and modified borderline-group methods in a small-scale OSCE. *Academic Medicine*, 77(7), 729–732.
- Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). New York: Macmillan.
- Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64, 425–461.
- Kobayashi, M. (2002). Method effects on reading comprehension test performance: text organization and response format. *Language Testing*, 19(2), 193–220.
- Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.
- Linacre, J. M. (2006). *A user's guide to WINSTEPS/MINISTEP: Rasch-model computer programs*. Chicago, IL: Winsteps.com.
- Livingston, A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.
- McNamara, T. (1996). *Measuring second language performance*. New York: Longman.
- Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5–11.
- Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 14, 3–19.
- North, B. (2000). *The development of a common framework scale of language proficiency*. New York: Peter Lang.
- Pitoniak, M. J. (2003). Standard setting methods for complex licensure examinations. *Doctoral Dissertations Available from Proquest*. AAI3078711. <https://scholarworks.umass.edu/dissertations/AAI3078711>.
- Popham, W. J. (1978). As always, provocative. *Journal of Educational Measurement*, 15, 297–300.
- Ramsey, P. (1980). Exact type 1 error rates for robustness of student's t-test with unequal variances. *Journal of Educational Statistics*, 5(4), 337–349.
- Schneider, J.W. (2017). *Null hypothesis significance tests: A mix-up of two different theories—The basis for widespread confusion and numerous misinterpretations*. <https://arxiv.org/ftp/arxiv/papers/1402/1402.1089.pdf>. Accessed July 2017.
- Schulz, W., & Fraillon, J. (2011). *The analysis of measurement equivalence in international studies*.
- Shepard, L. (1984). Setting performance standards. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 169–198). Baltimore, MD: Johns Hopkins University Press.
- Stansfield, C. W., Karl, J., & Kenyon, D. M. (1990). *The Guam educators' test of english proficiency (GETEP)* (Final Project Report, Revised). Washington, D.C.: Center for Applied Linguistics.
- Törmäkangas, K. (2011). Advantages of the Rasch measurement model in analysing educational tests: An applicator's reflection. *Educational Research and Evaluation*, 17, 307–320.
- Turner, C. E., & Upshur, J. (2002). Rating scales derived from student samples: Effects of the scale marker and the student sample on scale content and student scores. *TESOL Quarterly*, 36(1), 49–70.

- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 263–287.
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Houndmills, UK: Palgrave Macmillan.
- Wendt, H., Bos, W., & Goy, M. (2011). On applications of Rasch models in international comparative large-scale assessments: A historical review. *Educational Research and Evaluation*, 17, 419–446.
- Whitehead, Paul. (1986). *Statistics 2*. London: Pitman.
- Wright, B. D. (1992). IRT in the 1990s: Which models work best? *Rasch Measurement Transactions*, 6, 196–200.
- Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, 16(4), 33–45.
- Wyse, A., & Babcock, B. (2017). An investigation of undefined cut scores with the Hofstee standard-setting method. *Educational Measurement*, 36(4), 28–34.
- Wyse, A., Bunch, M., Deville, C., & Viger, S. G. (2014). A body of work standard-setting method with construct maps. *Educational and Psychological Measurement*, 74(2), 236–262.
- Zeiky, M. J., & Perie, M. (2006). *A primer on setting cut scores on tests of educational achievement*. ETS: Princeton.
- Zimmerman, B. B., Sudweeks, R. R., Shelley, M. F., & Wood, B. (1990). *How to prepare better tests: Guidelines for university faculty*. Brigham Young University Testing Services and The Department for Instructional Science. Brigham Young University. <https://testing.byu.edu/handbooks/bettertests.pdf>.

David Coniam is Chair Professor and Head of Department of the Department of Curriculum and Instruction in the Faculty of Education and Human Development at The Education University of Hong Kong, where he is a teacher educator, working with teachers in Hong Kong primary and secondary schools. His main publication and research interests are in language assessment, language teaching methodology and computer-assisted language learning.

Peter Falvey is a teacher educator, formerly a Head of Department in the Faculty of Education, The University of Hong Kong. His main publication and research interests are in language assessment, first and second language writing methodology and text linguistics.

Part II

The LPATE Enhancement Courses in Hong Kong: The Case of The Chinese University of Hong Kong

Barley Mak and Yangyu Xiao

This part gives a full account of the LPATE training courses provided for in-service English teachers between 2001 and 2006, with a particular focus on the courses offered by the Chinese University of Hong Kong (CUHK). Part II consists of three chapters. Chapter 9 introduces the initiative for the training courses and outlines the structures of training courses. Chapter 10 focuses on the Reading and Listening modules—the two components assessed by analytic means. Chapter 11 focuses on the Speaking, Writing, and Classroom language assessment modules, the three components wholly or partially assessed by scales and descriptors.

Chapter 9

The LPATE Training Courses: An Initiative to Improve Teacher Language Proficiency



Barley Mak and Yangyu Xiao

Abstract This chapter serves as an introduction to the LPATE training courses initiated by the HKSAR Government to provide in-service teachers of English with developmental and proficiency programmes. The current chapter provides a theoretical foundation highlighting the necessity for teacher professional development through training courses. This chapter then outlines the LPATE training courses provided for in-service teachers in Hong Kong between 2001 and 2005. Various modules put on by different course providers and the details of information on these modules are provided, including purpose of the training courses, the number of trainees enrolled and the attainment rate across different institutions. The chapter ends with remarks from trainees to illustrate how the training courses were perceived by course takers.

Introduction

The Language Proficiency Assessment for Teachers of English (LPATE) is a test of the standards of English language proficiency for Hong Kong primary and secondary school English teachers, or those who wish to become teachers of English. The needs for a benchmark assessment have been introduced in Section I of the current book. The initiative to introduce the LPATE, however, consisted of more than merely setting up a benchmark assessment as the HKSAR Government provided multiple channels for English teachers to achieve the language proficiency requirement (LPR). LPATE training courses were one such important avenue in facilitating this requirement.

B. Mak (✉)

United College, The Chinese University of Hong Kong, Sha Tin, Hong Kong
e-mail: barleymak@cuhk.edu.hk

Y. Xiao

Department of Curriculum and Instruction, Faculty of Education and Human Development,
The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, Hong Kong
e-mail: shirleyxiaoyy@gmail.com

The LPATE training courses were an example of such an alternative channel for eligible teachers to meet the language proficiency requirement. Eligible teachers included teachers holding a permanent post in public sector schools or local private primary/secondary schools offering full curriculum, who were teaching English or would be deployed to teach English. Those eligible teachers who took the training courses and passed the exit tests were considered to have met the language proficiency requirement. Thus, the LPATE training courses were an alternative channel for in-service teachers before 2000 to reach the benchmark requirement, whereas all new teachers who joined the teaching profession as from the school year 2001/2002 had to meet the language proficiency requirement through either exemption or the LPATE assessment.

The rationale for setting up the development courses was, in part, due to the backlash from teachers and their union against the imposition of the LPATE on all language teachers and where the government wished to show that the LPATE was not a one-off, stand-alone set of assessments but a comprehensive set of teacher development initiatives, hence the programmes and the workshops.

The LPATE training courses provided a range of valuable and well-resourced courses that helped English teachers develop the language proficiency required for teaching—as a government response to alleviating concerns about potential failure in the LPATE. The training courses and exit tests were expected to be designed to ensure that the language proficiency standards achieved or certified by the exit tests in the training courses and language proficiency standards as laid out by the LPATE were compatible. The courses and exit tests went through careful moderation and a number of reviews under a team of moderation panels. Support from various resources was also provided by the then Education Department to help course providers with course design, including development workshops, identification and dissemination of good practice, and the provision of feedback from internal and external examiners.

The LPATE enhancement courses were offered by course providers in both Hong Kong and overseas. The course providers were English departments or language centres at seven tertiary institutions in Hong Kong, three overseas universities and the British Council. These courses offered full support and assistance to language teachers who needed to improve their language proficiency and meet the language proficiency requirement. Teachers were allowed to enrol in one or all of the modules in each programme—whichever best suited to their needs.

The current chapter starts with an overview of the various courses provided by different course providers. It then focuses mainly on the perspective of one course provider—The Chinese University of Hong Kong (CUHK). The chapter gives a detailed description of the five modules provided by CUHK and illustrates how these modules were intended to help English teachers develop the language competence expected of them in the LPATE.

Training Courses for Language Teachers: Background

Importance for English Language Proficiency

English language teaching is a career that requires proficiency, pedagogy and professionalism. Recent studies also show English language teachers in Hong Kong becoming increasingly aware that English language teaching is a career that requires English language proficiency, subject matter knowledge and pedagogical knowledge (Coniam, Falvey, & Xiao, 2017; also Chap. 18, this volume). To help second language learners develop their language skills, English language teachers must have adequate and appropriate language proficiency and confidence in teaching English to students (Andrews, 2003). English language teachers need to provide comprehensible target language input in communicative language teaching, so as to facilitate teaching English through English in classrooms (Pennington & Hoekje, 2010).

In Hong Kong, it was the case in the 1990s that English teachers, in particular those in lower band (low student ability) schools, taught English to children through Cantonese (Mak & White, 1997). In addition, in the light of the small numbers of English teachers who were professionally trained with subject matter knowledge, there were worries about low English standards in the local business sector (Tsui, Coniam, Sengupta, & Wu, 1994). In response to such a situation, the LPATE was introduced to assess the language proficiency of English teachers in Hong Kong. It was made a requirement in Hong Kong that after 1997, all primary and secondary teachers should be degree holders and be professionally trained (Coniam & Falvey, 2013; Lai & Grossman, 2008).

A number of parallels can be cited in the Asian context. In Japan, where the language proficiency of English teachers has also been a cause for some concern, the EIKEN (Jitsuyo Eigo Gino Kentei—*Test in Practical English Proficiency* in English) has been used as a benchmark test for teachers teaching English at school level. English teachers are expected to obtain Grade Pre-1 (the EIKEN has seven levels: from Grade 1—the highest level down to Grade 5—the lowest level) (Eiken Tests, 2017). In Australian state schools where English is the language of instruction, the English language proficiency of teachers has been considered to be central to the quality of education. The Professional English Assessment for Teachers (PEAT) which consists of four papers (Reading, Writing, Speaking and Listening) related to a classroom context is another example of a standards test to ensure teachers' English language standards (Murray, Riazi, & Cross, 2012). Although PEAT is not merely set for teachers of English, such a test supports the proposition that teachers need to have sufficient English language proficiency if they need to teach through English.

Training Courses for Teacher Professional Development

Teacher professional training has had a considerable positive impact on teachers' beliefs and behaviours (Borko, Jacobs, & Koellner, 2010). In the specific case of teacher professional development in English language teaching, research studies indicate that teachers with more teaching experience and professional training are more proficient in the language that is available for learners to learn, with a willingness to engage learners in grammatical related issues and an ability to anticipate learner difficulties (Andrews & McNeill, 2005).

While the LPATE was established to serve as a benchmark to ensure the standards of English language teachers, the HKSAR Government also provided upgrading courses to help English language teachers enhance their English language proficiency. Since the establishment of the Advisory Committee on Teacher Education and Qualifications (ACTEQ) in 1993, various institutional initiatives (such as the continuing professional development framework) were developed to encourage and facilitate teachers to engage in professional development activities (Mak, 2010).

Professional development can be achieved through formal and informal support activities and courses that are designed to help teachers develop as professionals (Coldwell, 2017). The research literature reveals that professional development courses exert impact on teachers' professional development. Harris' (2001) findings with a group of teachers in Hong Kong reveal that professional development is helpful when teachers reflect on their own practices; professional development and self-reflection enable them to apply what they have learned to their teaching practice.

Worldwide professional training courses for language proficiency have been provided to help teachers improve their language proficiency. Considering the specific needs of Hong Kong teachers, Mak (2013) suggests that taking a degree in English is likely to help students to master grammar and improve writing skills, and an extended period of experience living overseas enhances one's mastery of other aspects of English language such as idiomatic usage and slang. Pearson, Fonseca-Greber and Foell (2006) discuss the potential of improving language teachers' language proficiency through taking degree courses in American universities. They argue that—to provide more input and help learners maintain standards—upper division literature and culture courses offered to Years 3 and 4 students in American universities should be taught in the target language. Pearson et al. (2006) also suggest that the universities use institutional resources to support students' language development, by, for example, providing more exposure to the target language through study abroad programmes, immersion programmes within the institutions, or the chance for service learning and extracurricular language activities. Exposure to the target language is important for future language teachers. In Japan, it is also recommended that the professional development of language teachers needs to consider both target language competence and the language of classroom management (Igawa, 2013). In Australia, preparation courses for Professional English Assessment for Teachers Test (PEAT) are also provided for teachers from different disciplines and backgrounds, to help them improve their language proficiency and feel less anxious about the upcoming

Table 9.1 The LPATE course providers

Location	Course provider	Courses provided
Hong Kong	The Hong Kong Baptist University	Language proficiency course for teachers of English
	The Hong Kong Institute of Education—later renamed The Education University of Hong Kong (EdUHK)	Core professional upgrading for English language teachers
	The Hong Kong Polytechnic University	Language proficiency training course for serving English teachers
	The Hong Kong University of Science and Technology	Towards excellence: a language training course for in-service teachers of English in Hong Kong
	The City University of Hong Kong	Continuing education certificate in English language teaching
	Lingnan University, Hong Kong	English language enhancement course for school teachers
	The Chinese University of Hong Kong	Training programme for the English language proficiency certificate
Hong Kong	British Council	Foundation course in English language for teachers
Australia	University of Queensland	English for TESOL professional purposes
	Queensland University of Technology	Professional training courses for teachers of English

test (Murray et al., 2012). The examples above further support the assertion that while benchmark standards on language proficiency have been introduced in different places around the world, respective training courses have also been provided to help candidates achieve the required standards.

Overview of the LPATE Enhancement Courses

Course Providers

LPATE enhancement courses were offered from the 2000/01 school year until the 2005/06 school year in nine tertiary institutions in Hong Kong, Australia and New Zealand. A list of course providers and courses provided are summarised in Table 9.1.

As can be seen from Table 9.1, a range of courses was provided to cater for the different needs of English language teachers, with English teachers able to enrol in the courses that they felt best fit their own schedules and needs. The training

courses were intended to help English language teachers meet the language proficiency requirement set out by the Hong Kong Education Bureau in 2000.

Purpose of the Courses

Appendix “[Course Providers’ Course Descriptions](#)” presents the brief course descriptions provided by different course providers as stated in the circular memorandum (No. 562/2000) provided by the Hong Kong Education Bureau.

In general, the language enhancement courses, according to the courses providers, served the following purposes:

- To enhance the language proficiency of English teachers in the four skills area (Listening, Speaking, Reading and Writing)
- To support serving English teachers in further developing their language proficiency
- To enhance participants’ competence to function effectively as English language teachers
- To equip participants with skills which were readily transferable to the teaching context
- To enable teachers to use English competently in the classroom and in their professional interactions
- To promote participants’ language and professional skills mainly but not exclusively related to Hong Kong English classrooms
- To develop participants’ professionalism in language teaching practice.

The purposes mentioned above delivered a strong message that these courses would enhance English teachers’ language proficiency in the context of English language teaching and would hopefully extend beyond the confines of classroom English.

Modules Provided

The course providers offered a range of modules corresponding to the language requirements of the LPATE. A summary of the modules provided can be found in [Table 9.2](#).

Generally, the focuses of the training courses in both Hong Kong and overseas were restricted to fulfilling the syllabus requirements as stated in the LPATE requirement (Bridges, 2007). Thus, these courses covered the different aspects of the LPR which were assessed in the LPATE, namely Reading, Writing, Listening, Speaking and Classroom Language. The training courses provided English teachers with necessary and sufficient knowledge to meet the required assessment standards (Bridges,

Table 9.2 Modules provided by different course providers

Course provider	Modules provided					
	Reading	Writing	Listening	Speaking	Classroom Language	Others
The Hong Kong Baptist University	✓	✓	✓	✓	✓	
The Hong Kong Institute of Education (the now EdUHK)	✓	✓	✓	✓	✓	
The Hong Kong Polytechnic University	✓	✓	✓	✓	✓	
The Hong Kong University of Science and Technology	✓	✓	✓	✓	✓	Error analysis
The City University of Hong Kong	✓	✓	✓	✓	✓	
Lingnan University	✓	✓	✓	✓	✓	
The Chinese University of Hong Kong	✓	✓	✓	✓	✓	
British Council	✓	✓	✓	✓	✓	
University of Queensland	✓	✓	✓	✓	✓	Project portfolio independent learning or consultation
Queensland University of Technology	✓	✓	✓	✓	✓	Language experience in Australia

2007). Since course designers and lecturers adopted a communicative and learner-centred approach, the training courses were likely to support the development of English teachers' communicative competence along with a critical awareness of the professional language required for language classroom and interaction with other teachers (Bridges, 2007).

The modules provided by different course providers offered participants different choices. Participants who chose to take the modules offered by overseas institutions would have the chance to attend immersion programmes overseas. The immersion courses overseas were regarded as contributing to English teachers' speaking and listening skills as well as intercultural communication ability (Lockwood, 2015).

Of the ten participating institutions, nine offered courses on all five modules, except for the Hong Kong Baptist University, which did not offer the classroom language module. The courses offered at overseas universities provided participants with language experience through immersion. These modules were tailor-made to address the needs of language teaching in Hong Kong. The enhancement courses thus provided a range of opportunities for English language teachers to develop their language competence in a specific area according to their own needs, with English language teachers being permitted to enrol for one or several modules of the different training courses. Further, with a view to catering for the needs of in-service teachers, the training courses offered by the institutions in Hong Kong were offered on a part-time basis, ranging from 120 to 230 h—depending on the number of modules provided. The immersion courses at overseas institutions were offered on a full-time basis, ranging from six to eight weeks.

The LPATE Training Courses at the Chinese University of Hong Kong

The majority of the current section focuses on the LPATE training courses at The Chinese University of Hong Kong (CUHK) where full course materials and supporting documents are available.

The description and analysis of the LPATE courses at the CUHK are based on the course materials used between 2004 and 2005 for two reasons. First, the course materials between years 2004 and 2005 are the most complete and comprehensive sets of courses materials that are still available 15 years after the courses were introduced. Second, the LPATE courses offered at the CUHK were basically the same each year; hence, using the courses from a single year is representative of the other years.

Features of the CUHK Training Courses

As stated in the leaflet outlining the professional development courses for teachers put on by CUHK, the training programme had the following features:

- The programme was specially designed to meet the needs of the Hong Kong teachers.
- An interactive task-based approach was adopted.
- Topics and themes used in the training courses were relevant to English teachers' professional life.
- A supportive learning environment was provided with the security of continuous assessment.

The courses, in general, were designed to take English teachers up to, or even beyond, the minimum levels of language ability set in accordance with the language

proficiency requirement (LPR). The courses were designed to enhance English teachers' language abilities, enrich their teaching and improve their professional awareness and prospects.

Structures of the LPATE Training Courses

At the CUHK, modules were taught face-to-face and involved individual and group work in a task-based approach. All course materials were based on educational themes and aimed to extend the four skills of reading, writing, speaking and listening, to build up language awareness and to promote reflection on the teaching and learning process. Participants who took these courses were considered to be benchmarked if they satisfied all the course requirements and passed the exit tests. English teachers who were benchmarked through taking the LPATE courses were exempted from the LPATE.

The structure of the courses is presented in Fig. 9.1.

As outlined in Fig. 9.1, the LPATE courses comprised of five modules. Participants were expected to undertake continuous assessments and exit tests in order to be benchmarked. The purposes and outlines of the courses are described below, followed by an explanation of how assessments were conducted.

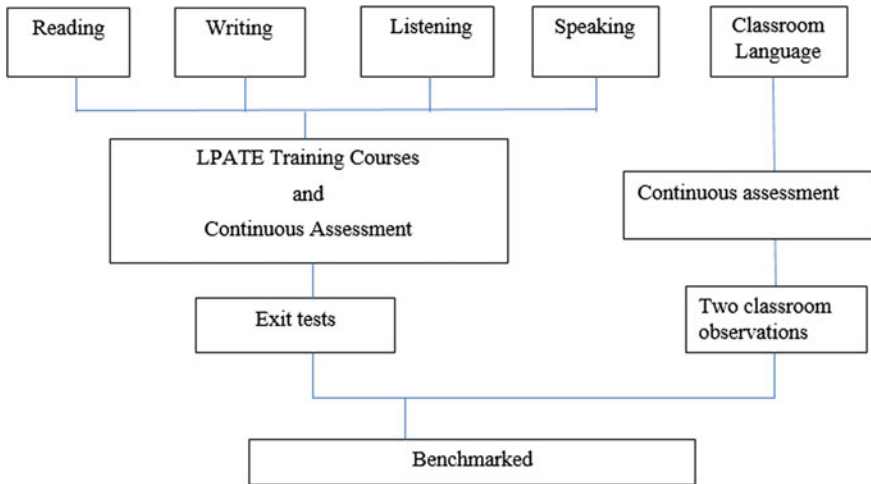


Fig. 9.1 Procedures for benchmarking

Reading Module

This module aimed to improve reading comprehension skills and extend participants' experience of reading. Passages on relevant themes from quality contemporary journalism and teaching methodology texts were used. Participants performed a variety of different tasks, involving different reading focuses and strategies.

Writing Module

This module aimed to improve writing skills and extend participants' knowledge of pedagogical grammar. Participants carried out a variety of writing tasks based on a school simulation. This module incorporated detailed consideration of principles and techniques of error correction, which was then extended into practical grammar and error correction exercises.

Speaking Module

This module covered the study of phonology and prosody, and the patterns of spoken discourse. Through detailed study of prose and poetic texts, participants were also encouraged to relate phrase structure and syntax to reading with meaning. Through a school-type simulation activity, participants learned to apply their developing language awareness in interaction with peers.

Listening Module

Programmes from TV and radio formed the basis of a study of phonological decoding, discourse structure and context- and genre-awareness, to build participants' confidence in listening to authentic texts. The broadcast programmes shared an educational theme and included discussions, debates, interviews, seminars and opinion pieces. Participants performed a variety of tasks based on their understanding of what they heard.

Classroom Language

Direct instruction, workshops, group work, role-play and micro-teaching were used to raise awareness of effective classroom language. Participants observed and prac-

tised dialogues relating to the language of the classroom. These included social and personal, organisational and instructional language. Participants were asked to record a sample of their own classroom discourse and use this text as the basis of a reflective exercise. Classroom study involved perceiving the link between teaching discourse and teaching methodology.

Continuous Assessments

Continuous assessments were designed to include a range of different activities and to cover a range of activities to test various language skills. Instructors followed a set of procedures for carrying out continuous assessment, which included giving prompt and meaningful feedback to participants, following the scales and descriptors of the LPATE. Where appropriate, participants were given copies of the instructors' feedback. Originals were kept by the instructors until the final grades were awarded.

Exit Tests

Exit tests closely modelled the LPATE. Exit tests were conducted according to the procedures of the LPATE. No examination drilling was carried out, and participants were not given prior warning of the content or themes of the exit test materials. The time allowed for answering exit test questions was the same as that allowed in the LPATE, except for the Listening Test. The reason why there was a difference lay in the fact that the Listening Tests at CUHK used authentic and unedited radio programme materials, and demanded a higher level of responses than in the LPATE itself.

In line with the practice at HKEAA, Speaking Test exit tests were assessed by two examiners. Writing Tests were marked by a third marker where there were discrepancies in results among different continuous assessment scores.

Marking Schemes

Participants were considered to be benchmarked if they satisfied the assessment requirement in the courses. For Reading, Writing, Speaking and Listening, continuous assessment grades counted for 50% of the overall final mark. Discretion was nonetheless applied. Where a participant had improved markedly during a course, for example, poor scores on the first assessment might be given less weight than assessments in the latter part of the programme. For the Writing module, where time did not permit all work to be done in class time, consideration was given to the fact that participants might have received help with work completed outside class—in gen-

Table 9.3 Course information by module

	Paper 1—Reading	Paper 2—Writing	Paper 3—Listening	Paper 4—Speaking	Paper 5—CLA
Total no. of classes	12	13	11	13	11
Total no. of participants who enrolled for the course	228	279	207	230	202
Total no. of participants who completed the course	225	268	197	224	192
Total no. of participants who passed	191	246	170	209	184
Overall attainment rates (%)	84.89	91.79	86.29	93.30	95.83

eral, at home, and these pieces were weighted somewhat less than those completed in classroom settings.

For Classroom Language Assessment, continuous assessment tasks were used formatively, and the overall grade was based on the assessment of the class teaching observed after the course.

The outlines above demonstrate that the LPATE enhancement courses were considerably more than an examination preparation course. The courses used a wide range of authentic tasks and exercises to help teachers achieve the expected language requirement and involved participants in reflection on their skills and practices.

Number of Participants and Attainment Rates

Table 9.3 provides an overview of the number of participants and attainment rates for the six-year period 2001 to 2006, as extracted from the final report submitted to EDB.

The attainment rates shown in Table 9.3 illustrate that the LPATE enhancement courses provided a good chance for trainees to develop their English proficiency up to the expected standards. The overall attainment rates indicate the percentage of participants who achieved a language proficiency level equal to Level 3 of the LPATE. Interestingly, Writing had a higher attainment rate than either Reading or Listening, in comparison with the consistently low attainment rate of the Writing Test in the LPATE (Coniam & Falvey, 2013; Lin, 2001; Appendix B “LPATE Results” in

Chap. 16 of this volume). It would appear that the training courses offered a good channel for participants to gradually improve their writing proficiency, especially in error correction and analysis—a major problem for many LPATE candidates—and to which considerable attention was paid in the LPATE Writing courses. The high pass rate across all modules appeared to support the conclusion that LPATE training courses were a good channel to enhance participants' language proficiency and help them meet the language proficiency requirement.

The LPATE Training Workshops

In addition to the LPATE enhancement courses, LPATE training workshops were also provided for PGDE students at the Chinese University of Hong Kong. These workshops intended to help participants develop a better understanding of the LPATE assessment so that the PGDE students had a greater chance of meeting the benchmark standards when they had to subsequently sit the LPATE. While the training programmes were only offered between the years 2001 and 2006, the workshops continued after 2006 until being finally discontinued in 2011.

The LPATE workshops consisted of six sessions:

Workshop 1: Overview

Workshop 2: Reading

Workshop 3: Writing

Workshop 4: Listening

Workshop 5: Speaking

Workshop 6: Classroom Language Assessment.

In each session, the course lecturer introduced the structure of the test paper, the major problems identified in the LPATE reports as reported by EDB (see http://www.edb.gov.hk/en/teacher/qualification-training-development/qualification/language-proficiency-requirement/lpat/lpat_assessment_reports.html for a list of the LPATE reports—accessed June 2017).

Focusing on the LPATE reports could inform teachers of the difficulties that candidates were likely to encounter when taking the LPATE assessment.

Course Evaluation and Feedback

Whereas detailed feedback forms from participants or course providers were, unfortunately, no longer accessible after fifteen years, the current section exemplifies participants' feedback from two available resources: (1) figures as obtained in the 2001–2002 course evaluation and (2) participants' feedback as submitted to the Educational Bureau in the concluding report. The participants were asked to respond to

a series of questions and score them on a six-point Likert scale, from strongly agree (6) to strongly disagree (1). The means of their responses are reported in Table 9.4.

Table 9.4 shows that generally the LPATE training courses were evaluated positively by course participants across all aspects, including content of teaching, assessment methods, and communication between teachers and students. Most items scored above 4, diverging considerably from the mid-point of 3.5, thus indicating strong acceptance of the proposition, i.e. that respondents wholly accept the argument (Bradshaw, 1990; Coniam & Falvey, 2013). In the current study, a “6” indicated a positive and “1” a negative response; thus, items which scored above 4 were considered to be strong positive responses. There were no items scoring below 3.5, indicating that participants generally held positive responses towards the training courses. Positive responses to items 1, 2, 3, 12, 13 and 14 indicated that lecturers who provided the training courses provided needed support and taught effectively in the classroom. Positive responses to items 4, 5, 15 and 16 showed that the training courses were designed at a level appropriate to participants’ level and were helpful in improving their language proficiency.

Scores of five items related to the Reading module (items 6–11) and six items related to the Listening module (items 5–6; items 7–11) were below 4 but above 3.5. It is possible that the authentic listening materials used in CUHK were challenging and participants found it difficult to apply to their own teaching. It could also be the case that since Reading and Listening are receptive skills, it takes time to transfer such skills to skills that are teachable to students.

However, the generally positive attitudes towards the training courses indicate that the training courses helped teachers with enhancing their English language proficiency so that they could meet the required LPR.

The responses from course participants were echoed by written comments provided by participants in the course evaluation forms. The comments in Table 9.5 were obtained from the final report submitted to the Educational Bureau dated 8 September 2006. This was when the whole training programme came to an end, after which all English teachers needed to meet the language requirement through taking the LPATE or by the exemption. LPATE training courses were no longer an alternative path for English teachers to meet the LPR, whereas LPATE workshops were still provided to help teachers to improve their language standards to meet LPR.

The comments above provide specific examples of how the training courses helped English teachers improve their language proficiency. The comments on the Writing Test, in particular, show that the training courses tackled the difficulties candidates have in taking the LPATE—correcting and explaining errors to students. Such comments may support the high attainment rate in the Writing Test, as shown in Table 9.3. The positive participants’ responses thus demonstrated how the training courses helped English teachers develop the language proficiency required for teaching.

Table 9.4 Course information by module (2001/2002)

Question	Reading	Writing	Listening	Speaking	CLA
1. I think she/he is genuinely committed to teaching	4.92	5.07	5.03	5.66	5.50
2. Through her/his teaching, I have come to understand this subject	4.44	4.82	4.62	5.46	5.23
3. She/he was able to explain difficult concepts clearly	4.66	4.94	4.87	5.45	5.25
4. My interest in this subject has been enhanced by his/her method of teaching	4.15	4.54	4.70	5.58	5.06
5. I thought the subject was pitched at a level suitable for their students	4.16	4.31	3.77	4.93	5.03
6. I thought the subject was covered comprehensively	4.00	4.03	3.96	5.00	4.80
7. I found the subject matter stimulating	3.99	4.26	4.10	5.06	4.96
8. I found the subject matter interesting	3.87	4.01	3.80	5.11	4.91
9. I think the content will be very useful for my future education	3.66	4.26	3.62	5.33	5.25
10. I found that my own understanding of my specialisation has been enhanced by undertaking the course	3.72	4.32	3.91	4.99	5.04
11. I thought the right amount of reference material was recommended	3.96	4.17	3.80	4.79	4.73
12. I thought the teacher made good use of the instructional equipment provided by the faculty/university	4.20	4.36	4.36	5.00	5.20
13. I think the teachers were sensitive to students' responses in class	4.73	4.87	4.76	5.64	5.47
14. I was able to communicate effectively with the teacher in the learning process	4.37	4.76	4.76	5.45	5.39
15. I thought the assessment methods of the course assignment/examinations were fair and appropriate	4.03	4.27	4.27	5.19	4.96
16. I thought the amount of work required for assessment was reasonable	4.10	4.40	4.40	5.23	4.55
17. As an overall evaluation, I would rate the teacher's performance as excellent	4.41	4.85	4.85	5.54	5.17
18. I would recommend this subject to others	4.03	4.43	4.43	5.28	5.07
Overall mean	4.19	4.48	4.48	5.26	5.08

Table 9.5 Participant comments identified in the final report

Modules	Comments	Suggestions
Reading	The course enhanced students' teaching skills The level of materials was found to be appropriate Direct quote from a student—"Yes, the course has inspired me"	More guidance for the final examination was requested
Writing	The examples on error correction were very helpful The discussions on writing assignments were found to be appropriate Direct quotes from students "The course has enhanced my knowledge of grammar" "I am able to explain most of the errors made by my students"	It would be better if more examples on essays could be given
Listening	The materials were found to be appropriate The course contained many useful materials The ongoing assessments were found to be suitable The course has enhanced the students' listening skills Direct quote from a student—"The course helped me convey the contents more effectively"	It would be more encouraging if the speakers in the videos can slow down the speed (pace)
Speaking	On teaching materials/contents The courses contained many useful and stimulating materials The level of materials was found to be suitable The lesson on phonology is interesting The materials learnt from this course Direct quote from a student—"If I know more, I can teach my students more information"	It would be better if the duration can be extended
CLA	On teaching materials/content Direct quote from a student—"My skill of eliciting has improved"	

The course providers felt that the courses were well organised and the students had shown interest in learning different topics. As the course providers reflected, the atmosphere in the workshops was also considered to be satisfactory. As one student noted, "The organisation and contents of the workshops were excellent and appropriate. I found the notes concise and helpful. I also appreciated the suggestions and encouragement from the course director".

It should be noted that some suggestions (e.g. providing more detailed guidance on the reading examination, adjusting the speed of an authentic video to suit students' individual needs) were not realistic. The CUHK course providers, did, however, try to follow up and integrate participants' suggestions into their subsequent course design.

Summary and Conclusion

The LPATE training courses were an integral component of an initiative to improve English language teachers' proficiency and help them to meet the LPR set out in the LPATE regulations. This chapter has provided an overview of the courses provided across different course providers in Hong Kong. The purposes of the courses, as stated by the course providers, together with teacher feedback, demonstrate that the LPATE training courses were professional development courses targeted specifically at the English language proficiency needed by English language teachers.

The following two chapters—Chaps. 10 and 11—provide a detailed account of the five LPATE modules and explain how these modules were able to contribute to English teachers' language proficiency.

Appendix: Course Providers' Course Descriptions

Course provider	Course description
The Hong Kong Baptist University	This course aims to provide opportunities for participants to enhance, in each of the four skills areas, the competence that they need in order to function most effectively as teachers of English. It will be offered on a part-time basis, over a period of 15 weeks, in evening and (for some modules) Saturday mornings. In Speaking and Writing, applicants may choose between two alternative modules; these cover the same syllabus but the modules with more contact time provide additional input and further practice for teachers who feel that they would benefit from it. Applicants may choose to take individual modules
The Hong Kong Institute of Education	This programme is designed to support serving English language teachers in further developing their proficiency in the use of professional and academic English by taking them to recognised level of competence. This is achieved through in-depth focus on the characteristics and use of the four fundamental language skills in the context of English language teaching. The programme is modular in design comprising and integrated professional upgrading. It builds on the foundations of teachers' professional knowledge and practical skills, enriching their knowledge base and introducing them to further applications hereby enhancing their professionalism as fully qualified English language educators
The Hong Kong Polytechnic University	The modular offering at the Centre for Professional and Business English is designed to target the individual skills of Listening, Speaking, Reading, Writing and Classroom Language Assessment. They are designed to be flexible, so that teachers can fit them in around their busy timetables. The individual modules are thematically based around the common educational concerns of Hong Kong teachers and authentic materials have been drawn from sources such as the Hong Kong media and educational journal articles. The language skills development work is task-based, and it is hope that teachers will find the modules relevant, enjoyable and helpful in their work as English language teachers in Hong Kong schools

(continued)

(continued)

Course provider	Course description
The Hong Kong University of Science and Technology	The course aims at enhancing participants' proficiency in English with reference to their specific needs as teachers of English. An important focus of the course is on equipping participants with skills which are readily transferable to their teaching contexts. Through face-to-face training and online learning, participants will have opportunities to deepen their understanding of key concepts, master essential skills, and reflect on their professional practice in the light of evaluation and feedback by themselves, their peer and experts in the related language areas. The synergy between face-to-face training and distance learning will further raise the effectiveness of learning as well as establish strong rapport between trainer and participants and among participants
The City University of Hong Kong	The course is designed for in-service secondary school teachers of English and comprises five modules (Listening, Speaking, Reading, Writing, Classroom Language) which can be taken as a package or individually. The courses aims at promote participants' language and professional skills mainly but not exclusively with reference to the Hong Kong classroom. Modules are designed and delivered by sympathetic, qualified tutors who have local and international teacher training experience
Lingnan University	This is a tailor-made English Language enhancement course for serving primary and secondary teachers of English. Participants will also be assisted to meet the language proficiency requirement set by the HKSAR Government. The objectives of the courses are for participants to: (a) develop English language oral and written communication skills and gain confidence in using English in and outside classroom; (b) engage in ongoing language skills enhancement; (c) have models for and actively develop professionalism in language teaching practice; and (d) meet or exceed the basic language proficiency level for teaching English in primary and secondary classrooms There will be an emphasis on expert instruction, reflection, self-development and communicative activities in a relaxed and enjoyable environment with maximum flexibility. The course materials have been specifically tailor-made and will include topical Hong Kong issues
The Chinese University of Hong Kong	This programme is designed to develop the English language skills of teachers of English in Hong Kong in order that participants can meet, and go beyond, the minimum acceptable level of language abilities as determined by the Advisory Committee on Teacher Education and Qualifications This programme will provide instruction in the essential language skills which teachers of English require in order to be effective in the classroom and in their general educational setting. While raising levels of language proficiency to meet the established standards, it will also equip the participants with subject content knowledge and enhanced language awareness, thus enriching the professional and personal development of English language teachers

(continued)

(continued)

Course provider	Course description
The British Council	The English Language Enhancement Course for Teachers (ELECT) consists of five free-standing modules that cover the requirement of Hong Kong Government's Language Proficiency Assessment for Teachers. The methodology is communicative and task-based. The modules are for teachers already close to the language proficiency requirement. Assessment is through three pieces of work, spread throughout the period of study. The exit assessment is over 50% of the overall assessment. Successful candidates meet official government requirement
University of Queensland	The English for TESOL professional purpose course is designed to enhance the English language proficiency and accuracy skills of teachers, while at the same time offering insights into creative and innovative methodology appropriate for teaching English to students in the Hong Kong school system The course, with classes delivered in Hong Kong and Australia, provides opportunities for participating teachers to build up their motivation and confidence to enable them to completely use English in the classroom and in their professional interactions. It will also focus on the intensive development of English language communicative competence and fluency in the four macro-skills of speaking, listening, reading and writing
Queensland University of Technology	The TESOL Unit in the University has an established record of excellence in second language teacher training and research with many years' experience in training teachers from all over the world, including many from Hong Kong. The 8-week full-time course offers participants the opportunities to go beyond the language proficiency requirement. The course will allow participants to broaden their language skills, to deepen their knowledge of the cultural understandings and contexts relevant to English language teaching and to enjoy the enriching experience of life both at an Australia university and with Australia family

References

- Andrews, S. (2003). Teacher language awareness and the professional knowledge base of the L2 teacher. *Language Awareness*, 12(2), 81–95. <https://doi.org/10.1080/09658410308667068>.
- Andrews, S., & McNeill, A. (2005). Knowledge about language and the 'Good Language Teacher'. In N. Bartels (Ed.), *Applied linguistics and language teacher education* (Vol. 4, pp. 159–178). US: Springer.
- Borko, H., Jacobs, J., & Koellner, K. (2010). Contemporary approaches to teacher professional development. In E. Baker, B. McGaw, & P. Peterson (Eds.), *Third international encyclopaedia of education* (pp. 548–556). Amsterdam, The Netherlands: Elsevier.
- Bradshaw, J. (1990). Test-takers' reactions to a placement test. *Language Testing*, 7(1), 13–30. <https://doi.org/10.1177/026553229000700103>.
- Bridges, S. (2007). Learner perceptions of a professional development immersion course. *Prospect*, 22(2), 39–60.
- Coldwell, M. (2017). Exploring the influence of professional development on teacher careers: A path model approach. *Teaching and Teacher Education*, 61, 189–198. <https://doi.org/10.1016/j.tate.2016.10.015>.
- Coniam, D., & Falvey, P. (2013). Ten years on: The Hong Kong language proficiency assessment for teachers of English (LPATE). *Language Testing*, 30(1), 147–155. <https://doi.org/10.1177/0265532212459485>.

- Coniam, D., Falvey, P., & Xiao, Y. (2017). An investigation of the impact on Hong Kong's English language teaching profession of the Language Proficiency Assessment for Teachers of English (LPATE). *RELC Journal*, 48(1), 115–133.
- Eiken Tests. (2017). <http://stepeiken.org/overview-eiken-tests>. Accessed September 2017.
- Harris, B. (2001). Facing the challenges of education reform in Hong Kong: An experiential approach to teacher development. *Pastoral Care in Education*, 19(2), 21–31. <https://doi.org/10.1111/1468-0122.00195>.
- Igawa, K. (2013). Language proficiency development needs of NNS English teachers in Japan. *Shitemouji University Kiyō*, 56, 191–216.
- Lai, K. C., & Grossman, D. (2008). Alternate routes in initial teacher education: a critical review of the research and policy implications for Hong Kong. *Journal of Education for Teaching*, 34(4), 261–275. <https://doi.org/10.1080/02607470802401370>.
- Lin, A. (2001). English language proficiency assessment for English language. *Reading*, 86(55), 63.
- Lockwood, J. (2015). The English immersion programme: Measuring the communication outcomes. *Indonesian EFL Journal*, 1(1), 107–116.
- LPATE Reports. (2017). http://www.edb.gov.hk/en/teacher/qualification-training-development/qualification/language-proficiency-requirement/lpat/lpat_assessment_reports. Accessed June 2017.
- Mak, B. (2010). The professional development needs of Hong Kong ESL teachers. *Asia Pacific Education Review*, 11(3), 397–410. <https://doi.org/10.1007/s12564-010-9073-5>.
- Mak, B. (2013). An investigation into the language needs of pre-service teachers of English for the language proficiency assessment for teachers (English) in Hong Kong. *Cypriot Journal of Educational Service*, 8(4), 381–390.
- Mak, B., & White, C. (1997). Communication apprehension of Chinese ESL students. *Hong Kong Journal of Applied Linguistics*, 2(1), 81–95.
- Murray, J. C., Riazi, A. M., & Cross, J. L. (2012). Test candidates' attitudes and their relationship to demographic and experiential variables: The case of overseas trained teachers in NSW Australia. *Language Testing*, 29(4), 577–595. <https://doi.org/10.1177/0265532212440690>.
- Pearson, L., Fonseca-Greber, B., & Foell, K. (2006). Advanced proficiency for foreign language teacher candidates: What can we do to help them achieve this goal? *Foreign Language Annals*, 39(3), 507–519. <https://doi.org/10.1111/j.1944-9720.2006.tb02902.x>.
- Pennington, M. C., & Hoekje, B. J. (2010). Language program as ecology: A perspective for leadership. *RELC Journal*, 41(3), 213–228. <https://doi.org/10.1177/0033688210380556>.
- Tsui, A. B. M., Coniam, D., Sengupta, S., & Wu, K. Y. (1994). Computer-mediated communication and teacher education: The case of TELENEX. In N. Bird, P. Falvey, A. B. M. Tsui, & A. McNeill (Eds.), *Language and learning* (pp. 352–369). Hong Kong: Government Printer.

Barley Mak is Associate College Head of United College at The Chinese University of Hong Kong. She is a teacher educator, working with primary and secondary English language teachers at the undergraduate and postgraduate levels. Her publications have appeared in a considerable number of internationally referred journals. She was the founding Director of the Centre for Enhancing English Learning and Teaching (CEELT), has conducted various public-funded research projects, and has served on a number of prominent HKSAR teacher education committees.

Yangyu Xiao is a Senior Research Assistant in the Department of Curriculum and Instruction at the Education University of Hong Kong. Her publication and research interests are in the fields of formative assessment, language curriculum and teacher education.

Chapter 10

The CUHK LPATE Training Courses: Reading and Listening



Barley Mak and Yangyu Xiao

Abstract This chapter introduces the Reading and Listening modules of LPATE training courses provided at the Chinese University of Hong Kong. The two modules are introduced together in one chapter as both Reading and Listening are assessed in an analytic manner in the LPATE. The chapter first introduces the two key aspects assessed in the LPATE Reading and Listening Tests, i.e. cognitive abilities as well as linguistic skills and knowledge. It then illustrates how the Reading and Listening modules helped participants achieve the expected language standards by giving examples of tasks used in the modules, supplemented with the course providers' interpretations.

Introduction

In the LPATE, both Reading and Listening Tests are analytically marked. The Reading and Listening courses are presented in the current chapter. The test items in the Reading and Listening Tests comprise relatively discrete items, which are used to assess two main skills or abilities: cognitive abilities, as well as linguistic skills and knowledge. Thus, a major function of this chapter is to discuss how the development of the above two aspects in reading and listening are supported in the LPATE training courses.

B. Mak (✉)

United College, The Chinese University of Hong Kong, Sha Tin, Hong Kong
e-mail: barleymak@cuhk.edu.hk

Y. Xiao

Department of Curriculum and Instruction, Faculty of Education & Human Development,
The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, Hong Kong
e-mail: shirleyxiaoyy@gmail.com

Listening and Reading Comprehension: Background

Existing research literature has discussed the connection between listening and reading skills and how the two skills are related (Lund, 1991; Mecartty, 2000; Park, 2004). The proponents who assert that there are connections between the two skills believe that both listening and reading require competence in decoding and comprehension (Lund, 1991). Decoding is a process to convert audio or written text into basic language units, whereas comprehension is a process of combing decoded text with the listeners' or readers' prior knowledge (Lund, 1991). Both listening and reading comprehension requires the audience or readers to decode audio or visual input, with the assistance of their prior knowledge.

It is natural to expect that reading and listening comprehension require linguistic and lexical knowledge (Mecartty, 2000). Mecartty's (2000) study of Spanish fourth-grade students reveals both lexical and grammatical knowledge being significantly correlated with listening and reading comprehension. A complex interaction between lexical and grammatical knowledge and comprehension has been observed, in that lexical and grammatical knowledge contribute different amounts to the total variance in comprehension. The study also reveals that good lexical knowledge contributes more to comprehension than grammatical knowledge. A further study that supports the role of grammatical and lexical knowledge is Park's (2004) study of college English learners in Korea. On the basis of the test performance in tests of linguistic skills, and reading and listening comprehension tests, Park (2004) reports the significant effect that linguistic competence exerts on L2 listening and reading comprehension. Park's (2004) interpretation of such a finding supports the proposition that both listening and reading comprehension have been realised through interpreting and parsing oral and written text. Such comprehension achieved in a bottom-up way is nevertheless important in comprehension.

Apart from linguistic knowledge and skills, cognitive abilities are also central to effective comprehension. Yamashita (2015) points out that readers need to have the cognitive ability to comprehend texts through using the following strategies: skimming and scanning to speed up the reading process; remembering selected information in the text; comprehending difficult texts by slowing down the reading process, or repetitively reading these words and phrases, or backtracking to the previous part of the texts. To facilitate the cognitive process of understanding reading passages, readers need to have relevant and basic linguistic knowledge and skills, as well as relevant background knowledge with the latter facilitating comprehension in a top-down manner (Park, 2004). Whereas cognitive abilities, and linguistic skills and knowledge are likely to have different impacts on reading and listening comprehension, as the review above indicates, they are two interconnected aspects that are needed for effective reading and comprehension. Accordingly, these two aspects are two key aspects assessed in the LPATE.

Table 10.1 Cognitive abilities in reading and listening

<i>Cognitive abilities</i>	
Local processing	The ability to retrieve information from phrases and sentences, such as recognition of numbers, date and time, facts, lexical terms and syntactic structure
Global processing	The ability to obtain the gist of a longer stretch of language such as a paragraph in a passage, or the language in a speaker's turn in a conversation
Inferencing	Predicting or inferring implicit meanings from lexical items, cohesive devices or colloquial expressions
Interpreting language in a larger context	The ability to interpret language with an understanding of its larger linguistic context, the physical setting or the social situation
<i>Linguistic skills and knowledge</i>	
Conceptual meaning	The meaning of a word or a group of words—such as words and phrases, idioms and colloquial expressions
Propositional meaning	Meaning at the sentence level
Textual and rhetorical meaning	The arrangement of propositions which form a coherently structured text. Rhetorical functions include asserting, introducing, justifying and clarifying, etc.
Pragmatic meaning	The interactional aspect of communication, for example, the writer's or speaker's intention, action, stance, etc.

Assessment of Reading and Listening in the LPATE

The abilities which the LPATE Reading and Listening tests tap may be seen from two perspectives: *cognitive abilities* and *linguistic skills and knowledge*. *Cognitive abilities* include the abilities to process reading texts from four different aspects: local processing, global processing, inferencing and interpreting language from a larger context. *Linguistic skills and knowledge* are needed to extract meanings from written or aural texts at four levels: conceptual meaning, propositional meaning, textual or rhetorical meaning and pragmatic meaning. Reading comprehension is a process whereby readers need to interact with the written texts using reading skills and their prior knowledge.

Table 10.1 describes and summarises the main components of cognitive abilities and linguistic skills and knowledge, on the basis of the information retrieved from the LPATE handbook (Government of the Hong Kong Special Administrative Region, 2000, p. 5).

The Table 10.1 illustrates how the LPATE attempts to assess candidates' reading and writing abilities by drawing attention to similar abilities and knowledge required for reading and listening comprehension.

The Reading Module

The Structure of the Reading Module

The LPATE Reading courses were designed in such a way as to enhance participants' cognitive abilities and linguistic skills and knowledge as assessed in the LPATE. The Reading module comprised four units: (1) reading skills; (2) reading processes; (3) reading for pleasure; and (4) textual organisation, as Table 10.2 outlines.

The following subsections will draw attention to how the reading tasks laid out in Table 10.2 support the development of the cognitive abilities and linguistic competence expected in the LPATE assessment.

Reading Skills

As stated in Table 10.2, five aspects of reading skills were practised in the reading module: (1) identifying view points; (2) understanding the structure of a text; (3) identifying facts and opinions; (4) textual cohesion and (5) guessing meaning from context. The section below exemplifies how the relevant reading tasks were used to practice respective reading skills using selected examples. An interpretation of how these tasks contribute to the development of cognitive abilities, linguistic skills and knowledge is also provided.

Table 10.2 An overview of the focus of the reading module

Units	Major focuses of each unit	Relevant tasks
Reading skills	Identifying viewpoints	Tasks 1, 2, 3 and 4
	Understanding the structure of a text	Tasks 5 and 6
	Identifying facts and opinions	Tasks 7, 8 and 9
	Textual cohesion	Tasks 10, 11 and 12
	Guessing meaning from context	Task 13
Reading processes	Understanding the secrets of reading	Tasks on reflecting on secrets of reading
	Decoding syntactic structures	Tasks 14
	Identifying top-down and bottom-up concepts in a text	Tasks 15
Reading for pleasure	Understanding literal and metaphorical language	Task 16
	Skimming book reports to provide guidance for students	Task 17
Textual organisation	Understanding how lexical families help achieve textual cohesion	Tasks 18 and 19
	Understanding different ways of textual organisation	Tasks 20, 21, 22 and 23

Table 10.3 Identifying viewpoints

Reading passages	Reading an article “How come when our students graduate, they can’t compete and do their jobs properly” and discuss the viewpoints put forward in the article
Task 1 Identifying different people’s viewpoints	
	Do you agree with the opinions put forward by the senior educators and civil servants? Whose viewpoint is being presented here? Is it the opinion of the writer or that of the senior educators or civil servants?
Task 2 Understanding writer attitude	
	Analyse how negative attitudes are communicated by the writer Find where the writer draws an analogy. Explain the analogy Identify the speech of Tai Hay Lap on items of vocabulary that are heavily negative in their connotation
Task 3 Discussing readers’ own views	
	Do you agree with the following statement? This is an example of a piece of journalism which presents a negative view of the Hong Kong education system; the journalist does not give any personal opinions, but presents highly critical statements of those involved in the discussion
Task 4 Interpreting writer tone	
	<i>Senior educators, civil servants and executive councillors on the commission took turns to denounce the elitist and examination based system <u>that made their careers</u></i> What is the function of the underlined phrases? How would you describe the writers’ tone here?

Identifying Viewpoints

The four tasks on identifying viewpoints offered participants opportunities to interpret language in both a larger context (Task 1 and 3) and a local context (Tasks 2 and 4). In the larger context, Task 1 required participants to develop an understanding of the viewpoints of different people whereas Task 3 required participants to discuss their own views. In the local context, Tasks 2 and 4 expected participants to identify and understand an analogy or a specific phrase used in the text.

In terms of linguistic skills and knowledge, the tasks tapped pragmatic meaning in texts (Tasks 1 and 3), as well as the conceptual meaning of words and phrases (Tasks 2 and 4). On the whole, the tasks attempted to tap different aspects of cognitive abilities and linguistic skills specified in the LPATE handbook (Government of the Hong Kong Special Administration, 2000) (Table 10.3).

Table 10.4 Understanding the structure of a text

Reading passages	Reading an article about Hong Kong employers' employment preferences (based on a survey conducted by the Open University of Hong Kong) in the South China Morning Post and complete the tasks below	
Task 5 Dividing the text into sections		
	If you were to divide this text into sections, how many parts do you think it would have? Suggest sub-headings for the sections	
Task 6 Identifying the advantages and disadvantages of different approaches to learning		
	In groups fill in the following table according to the information given in the text	
	Advantages	Disadvantage
Web-based learning		
Lecturer-based learning		

Understanding the Structure of a Text

Two tasks on understanding the structure of a text expected participants to become aware of how information is organised and structured in a reading text (see Table 10.4). More specifically, Task 5 expected participants to work out how propositions were arranged coherently in a structured text; Task 6 required participants to identify two key aspects in the reading text, i.e. advantages and disadvantages of web-based learning and lectured-based learning. Participants were likely to notice how key information was organised in a passage by identifying such information. Both tasks addressed the cognitive abilities at a global level; to complete the tasks, participants needed to understand the rhetorical and pragmatic meanings of the language.

Identifying Fact and Opinion

Table 10.5 illustrates three tasks on identifying fact and opinion. In these tasks, participants were expected to distinguish what facts were and what opinions were (Task 7); they also needed to identify how evidence was used to support opinions (Task 8) and how writers express their opinions using hedging (Task 9).

The ability to distinguish fact from opinion indicates that readers can understand the true meaning of the authors as well as the hidden message behind it (Che, 2002). Such ability is important to work out the real purpose of a reading passage. To identify facts and opinions, participants need to understand the pragmatic meaning of a text, and how language, such as words and expressions, is used to indicate the purpose of the writer and the stance the writer holds.

Table 10.5 Identifying fact from opinion

Task 7 Distinguishing facts from opinions	
	In this text, identify some examples of FACTS and OPINIONS
Task 8 Using evidence to support opinions	
	Sometimes opinions are stated as facts because the speaker has evidence to back up his/her opinion. e.g., “If you combine Web-based and lecture-based instruction, students tend to do better” What is the evidence in the text to back up this statement?
Task 9 Identifying words to show attitudes	
	Find some examples from the text of tentative conclusions

Textual Cohesion

Textual cohesion is related to how words, phrases and sentences in a text are linked together to express meaning, as well as how ideas are made explicit in a text (Graesser, McNamara, Louwerse, & Cai, 2004). The abilities to understand how a text is coherently constructed facilitate the comprehension of the text (Ozuru, Dempsey, & McNamara, 2009). Tasks 10–12 addressed the issue of cohesion by drawing attention to how paragraphs are meaningfully connected together in a text (Table 10.6).

The three tasks made explicit how cohesion may be achieved through connecting sentences in a meaningful manner, in addition to other strategies, such as using cohesive devices. Tasks 10 and 11 were likely to urge participants to think about the connection, whereas Task 12 allowed participants the chance to practise developing coherence texts through rearranging jumbled sentences. In terms of cognitive abilities, these tasks addressed the processing of texts at a global level and interpreting language in a larger context, into an understanding of the whole structure of the text.

Table 10.6 Textual cohesion

Task 10 Understanding a passage from the opening paragraph	
	Read the opening paragraph from a newspaper article and decide what it is about
Task 11 Understanding how meaning is conveyed in a text	
	Read the next paragraph and predict how the writer weaves this piece of information about a 15th century painter into the article, in the second paragraph
Task 12 Connecting jumbled sentences	
	Read jumbled sentences from a text. Try to recreate the paragraph by rearranging these sentences

Guessing Meaning from Context

Guessing meaning from context is a global reading skill. In reading a text, it is natural that readers may encounter unfamiliar words whose meaning they need to interpret from the context. Guessing meaning from context is a metacognitive skill that readers can use to cope with a lack of vocabulary knowledge in reading by making inferences through contextual clues (Zhang, 2001). It is common for second language learners to meet unfamiliar words in reading. In comparison to checking the meaning of every word, guessing meaning from context is likely to facilitate and speed up the reading process (Zhang, 2001).

Task 13 provided a chance for participants to experience how meanings of words can be derived through making inferences (see Table 10.7). The task was likely to raise participants' awareness that they could cope with difficulties in reading through guessing and making inference.

To sum up, the unit on reading skills provided the opportunity for participants to practise, use and develop their reading skills. The selected tasks or task extracts above demonstrate how such reading skills were enhanced through reading tasks. These reading tasks, while developing participants' own reading skills, were also intended to raise awareness of how different reading skills may facilitate reading comprehension in their own students.

Reading Processes

The unit on *Reading Processes* was designed to draw attention to why and how readers comprehend texts. The unit started with raising participants' awareness of why and how people comprehend a text; after which the module focused more specifically on 'decoding syntactic structures' and 'top-down and bottom-up concepts in reading'.

Table 10.7 Guessing meaning from context

Task 13 Guessing the meaning of words from the context	
	<p>Example (by Barton, 2002, There is more to life, in <i>the Guardian</i>): Try to keep healthy. Don't stay up late into the night surviving on fizzy cola, endless cups of coffee and three packets of Maryland cookies as you wade through all those mathematical formulae What does 'wade through' mean: (a) Work systematically; (b) complete on schedule; (c) struggle to get through; (d) fail to finish</p>

The Secrets of Reading Comprehension

The session on reading processes started from a piece of awareness-raising text with the topic of ‘understanding the secrets of reading’. Participants were provided with three texts focusing on three secrets of reading, namely ‘the magic of print’, ‘the magic of language’ and ‘the magic of general knowledge’, which are the three key aspects of knowledge a reader should have in order to understand the text.

The text on ‘the magic of print’ demonstrates that it would be difficult for people to understand a series of symbols, such as ‘D.o.c)...f. J7toAoLay+!D.’; similarly, it would also be difficult for those who only know English to understand a Russian edition of the Old Testament. The text aimed to raise participants’ awareness that readers need to know symbols of the text they read in order to understand them.

The text on ‘the magic of language’ further demonstrated the truism that people need to understand the language if they are to read in that language. Thus, they need to know more than the letters of the language. This understanding includes understanding the meaning of words, the way in which words are put into sentences, paraphrases and passages. The most effective way to improve reading, according to the text, is to read more or to read more diverse and difficult texts.

The text on the ‘the magic of general knowledge’ informed participants that readers need relevant knowledge to understand a text. For example, a teacher who does not skateboard may well find such terms as *ollie*, *pop*, *fakie*, *grind*, *goofy* somewhat puzzling. Similarly, students need to read more and more diverse texts, in order to enrich their knowledge about general knowledge. The requirement in general knowledge echoes what has been mentioned in the literature review that readers need background knowledge to facilitate understanding (Park, 2004).

To sum up, this introductory session on the reading process aimed at making the reading process explicit to participants, through reading. The session drew participants’ attention to three key aspects they need to understand the reading passage: the written language as a set of symbols, the meaning of the language and relevant subject knowledge. The importance of this approach is that the participants benefited in two ways: first, they themselves improved and became more aware of reading skills; second, they became much more aware of what their students go through in the reading process and become exposed to the kinds of tasks that will help their students.

Decoding Syntactic Structures

Reading comprehension also requires decoding the syntactic structures (Hedge, 2003), which means that readers need to have linguistic knowledge of a language. A study of fifth- grade students in the USA revealed that children’s syntactic awareness is closely associated with their reading fluency ($r = .625$) and their reading comprehension performance ($r = .816$); thus, poor syntactic awareness corresponds

Table 10.8 Decoding syntactic structures

Task 14 Understanding the role of syntactic structures	
	<p>Read the following nonsense passage and then see how many of the questions you can answer. (Jane Willis's nonsense passage experiment, from Willis 1981, <i>Teaching English through English</i>, p. 150)</p> <p>"The grifty snolls cloppered raucingly along the unchoofed trake. They were clary, so they higgled on, separately. "Ah, chiwar kervay," they squopped rehoply. "Mi psar Quaj!" "Quaj!" snilled one, and filted even jucklier".</p> <p>(a) Where did the snolls clopper? (b) What was the trake like? (c) Why did they higgel on? (d) Why did they clopper raucingly? (e) What did they do separately? (f) Would an unchoofed trake be easy or difficult to drive a car on? (g) Did the snolls travel quietly or noisily? How do you know? (h) What was the name of the place they were going to?</p> <p>Discussion: Did you understand the passage? How many of the questions were you able to answer? Which questions? Discuss why you were able to answer some questions and not others? What has this nonsense exercise shown you about writing comprehension questions on reading passages?</p>

to poor reading fluency and poor reading comprehension (Mokhtari & Thompson, 2006). Task 14 in Table 10.8 illustrates the role of syntactic structures in reading comprehension.

The above task on the nonsense language text provided a chance for participants to think about why they can respond to some of the reading comprehension questions even though they do not have a complete understanding of the passage. English teachers who took the course realised that while they were able to answer the questions depending only on recognising the word classes and syntactic structures, they were not able to answer the questions assessing the actual meaning of words. The task also made participants aware how far they may understand a text by paying attention to syntactic structures.

Top-Down and Bottom-Up Concepts in Reading

Top-down and bottom-up reading are two aspects of metaphorical modes of reading which are related to how comprehension is carried out (Grabe, 2008). More specifically, the stereotypical discussion of top-down and bottom-up reading process regards reading letter by letter, word by word, and sentence by sentence, as the bottom-up reading process, whereas a top-down reading process is a process by which readers are clear about the reading goals, actively control the comprehension process and are able to look for relevant information to achieve the comprehension

Table 10.9 Identifying top-down and bottom-up concepts in a text

Task 15 Identifying top-down and bottom-up concepts in a text		
You are going to read a section of a chapter on reading skills from <i>Steven M. McDonough 1995 Strategy and skill in learning a foreign language</i> Your task is to complete a table, identifying various concepts from this text (as shown below) as being more associated with ‘top down’ or ‘bottom up processes’		
(a) Skillful decoding	(h) The use of language specific knowledge	
(b) Relating information to a reader’s prior knowledge	(i) The use of pre-existing knowledge of text structures	
(c) Text driven	(j) Predicting and anticipating events and meanings	
(d) Concept driven	(k) Recognition of syntactic structure	
(e) Rapid, context free word recognition	(l) Inference of meaning from a wider context	
(f) Higher level ‘guessing game’ strategies		
(g) Reciprocal perceptual/cognitive process		

goals (Grabe, 2008). However, no reader is a pure bottom-up reader or top-down reader. Effective readers are able to use both processes to facilitate comprehension. Table 10.9 illustrates the task related to top-down and bottom-up reading processes.

Task 15 was essentially an awareness-raising exercise. After reading the text, participants had the opportunity to reflect on the reading process and relate these different strategies to a top-down or a bottom-up approach.

Reading for Pleasure

Reading for pleasure is one of the most common purposes for reading. The strategies needed for reading for pleasure are different from reading for academic purposes or reading to answer test questions. The following sections demonstrate the process involved in reading for pleasure using two examples.

Literal and Metaphorical Language

Reading texts may contain both literature and metaphorical language. When doing reading comprehension, readers need to infer the metaphorical meaning of an expression from time to time. It is believe that second language learners who are more aware of metaphorical language are more likely to approach reading cognitively, affectively and pragmatically in a more effective way (Boers & Lindstromberg, 2006).

Participants in the reading module read a piece of contemporary journalism from the ‘features’ section of a newspaper. As this piece of news was in a section dealing with issues such as culture, arts, travel and lifestyle, it contained a number

Table 10.10 Literal and metaphorical language

Task 16 Understanding literal and metaphorical language	
Sample text (extract)	This list is not intended to be definitive. It is merely a jumping-off point, a place to start exploring the world of books. In recent years publishing for children has become a growth area. The shelves of bookshops, but not, alas, our cash-starved libraries-are stuffed with new titles and classics. Where to begin? How to choose? We hope that this list will help you and your children and teenagers plunge in and develop your own taste and own likes and dislikes (Selected from: Dickens to Dahl: the classic reads, in <i>the Guardian</i>)
Reading comprehension questions	Read the passage and try to explain the literal and metaphorical meanings of the following words and expressions: Cash-starved libraries Stuffed with new titles Plunge in

of metaphorical expressions. Table 10.10 presents a selected text and the related metaphorical expressions.

In the reading tasks, participants were expected to work out the meanings of the metaphorical phrases, drawing on the context of the reading passage for meaning. Such a skill is needed for reading stories, novels and literature.

Skimming Book Reports

Skimming helps readers pick out relevant and useful information. Participants were asked to skim-read a series of book reviews so that they might be able to provide guidance to students in their extra-curricular reading. Such an activity was authentic and mimicked the situation that English teachers might encounter in their classrooms. To complete the task, participants needed to process information at a global level and retrieve the most relevant information (Table 10.11).

With the assistance of the key information as listed in the table, participants were able to derive the information they need and practice the skill of skimming. Through doing this task, they would, it was hoped, become aware of how and where to quickly retrieve the information that they need.

Table 10.11 Skimming book reports to provide guidance for students

Task 17 Skimming book reports to provide guidance for students				
Skim-read a series of book reports and summarise the following information:				
Title	Genre	Themes	Age group	Who might it especially appeal to?

Textual Organisation

Textual organisation relates to how coherence in a reading passage is achieved at lexical, sentence and paragraph levels. Tasks 18 and 19 were on the topic of cohesion at lexical level (see Table 10.12). More specifically, the tasks were about how words with similar or different meanings were used together to achieved coherence. Task 18 demonstrated an example of similarity, and Task 19 presented an example of difference.

Tasks 18 and 19 drew attention to the issue of ‘lexical families’. A good command of such skills may help participants guess the meaning of the words from context if they are not familiar with the word meanings in a text.

Tasks 20–22 focused on textual cohesion at the sentence level and above (see Table 10.13). At the sentence level, cohesion is achieved through different ways of textual organisation, through using vocabulary with different meanings (e.g. from general to specific meanings or words indicating the logical sequence of sentences). Reading texts can be coherently connected through vocabulary indicating the different logic of the texts, which include: describing an event from general to specific (Task 20), describing an event in time sequence (Task 21), indicating the cause and effect relationship in a text (Task 22) and indicating the contrasting relationship between sentences (Task 23).

Table 10.12 Textual cohesion

Task 18 Lexical families: related meanings	
Reading comprehension questions	At the lexical level, writers often pair words and expressions with similar or related meanings, or even create groups of three similar expressions Complete the blanks below: ‘Disruptive’, ‘violent’ and ‘abusive’ are words used to describe children and _____ who are suspended or _____ from school. But aggression and _____ call for more than just _____ and exclusion List of words: abuse, expelled, young people, punishment
Task 19 Lexical families: similar or different	
Reading comprehension questions	Try to decide whether the missing words should be similar to or in contrast to the underlined word or expression in the text I don’t think smacking should be banned. It is a form of discipline. If there wasn’t smacking a lot more children would be _____. The government can’t choose to control how a parent brings up their child. It’s only when smacking turns into abuse that they should be concerned and step in and do something The missing word is similar/different The word could be _____ Answer: Contrast: out of control

Table 10.13 Different ways of achieving textual organisation

Task 20 From general to specific reference	
	<p><i>Analyse the two sentences below and give examples of 'general' and 'specific' references</i></p> <p>"I started getting involved with drugs when I was about 12. I began sniffing nail varnish at home and then started on amphetamines"</p> <p>"Then the school put us in touch with The Children's Society. Straight away they arranged for Ryan to see one of their workers. The change was almost immediate. He calmed down a lot and was more ready to control his temper. Now he counts to ten if he begins to feel angry, and it works"</p>
Task 21 Temporal sequence	
	<p><i>Underline all the words which show the time sequence</i></p> <p>I was eight when my stepdad started sexually abusing me. It went on for years before I had the courage to tell mum. Then she didn't believe me. I ran away from home time after time. I ended up in a children's home. Because I was the youngest there, the others bullied me. None of the adults seemed to care. In the end, I ran away from there too</p>
Task 22 Logical sequence—cause and effect	
	<p><i>Notice how, in this text, the events of the girl's life seem to have a clear cause-and-effect relationship</i></p> <p>The streets were the only place to go. I got to know other homeless people. They gave me drugs that made me forget everything that had happened to me. I also met men who offered me money in return for sex. What could I do? I had to live. I was even arrested and convicted for soliciting</p>
Task 23 Logical sequence—contrast	
	<p><i>Identify the turning point in the following paragraph</i></p> <p>The root of the problem was that Ryan felt he was doing everything wrong. He never received praise for anything he did. Working with the Children's Society changed that. He did the things he enjoyed and was good at. We also saw that we had to encourage him. This built up his confidence, and he realised he could be well behaved too</p>

Summary

In summary, a variety of tasks were provided in the Reading module to enhance participants' reading abilities in terms of both cognitive abilities and linguistic skills and knowledge. These included retrieving information at local and global level; understanding how information is organised and structured in a text; understanding the textual and rhetorical meaning of the text; distinguishing fact from opinion; guessing meaning from context; decoding syntactic structures; understanding literal and metaphorical meaning of the language and skimming to locate relevant information. The explicit focus on reading process was likely to raise participants' awareness of the reading process through reading.

The reading tasks used in the Reading module were related to education or were familiar to language teachers in Hong Kong. The reading tasks were authentic and

were likely to help participants to meet the language standards they needed for teaching in a Hong Kong context.

The Listening Module

The Structure of the Listening Module

In the LPATE Listening Test, candidates are expected to listen to texts of various types which include discussions, debates, interviews and documentaries. These text types are of various genres, accents and speed. It is noted that candidates need to understand both the factual details and to derive a deep meaning from the content. In response to the expectation of the LPATE, the training courses thus attended to both micro-listening skills and macro-listening skills, as well as providing practice on a range of text types.

The listening module consisted of seven sessions. Table 10.14 summarises the focus of each session and respective tasks used to address such a focus.

The following sessions elaborate how the LPATE listening modules supported participants' listening abilities using examples from the course.

Table 10.14 Overview of the sessions in the Listening module and listening tasks

Sessions	Focus of the sessions	Relevant tasks
Phonological, grammatical, syntactic and semantic skills	Foundation skills for listening	Tasks 1–7
Micro-skills in listening	Discriminating and identifying sounds and syllables	Tasks 8–22
Macro-skills in listening	Predicting, inferring, recognising cohesive device and following a longer stretch of text	Tasks 23–25
Following an argument	Recognising the structure of an academic talk and reflecting the process of listening to academic talk	Task 26
Understanding attitudes and opinions	How authors' attitudes and opinions are expressed and supported by relevant evidence	Tasks 27–28
Drawing comparisons	Extracting main issues and identifying key points of differences and similarity	Task 29
Current topics in education	Listening to issues related to education issues	Task 30

Foundation Skills for Listening

In order to understand a listening text, a listener needs to have basic language skills, which include phonological, grammatical, syntactic and semantic skills.

Phonological Skills

Phonological skills consist of a broad awareness of sounds, within which people can identify and manipulate small units of language, such as words, sounds and rhymes. In the LPATE training courses, English teachers practised a range of exercises to raise their phonological awareness. The English teachers listened to a series of sentences and were then asked to distinguish singular and plural nouns (Task 1); distinguish present and past tense (Task 2); distinguish whether the definite or indefinite article was used (Task 3). Selected test questions are presented in Table 10.15.

The three tasks above attended to the most fundamental competence in distinguishing words with similar sounds. Such tasks addressed cognitive abilities at local level and the linguistic competence in retrieving meaning at the level of words.

A further exercise on distinguishing sounds is presented in Table 10.16 where participants needed to use phonological as well as grammatical knowledge.

Table 10.15 Phonological awareness

Task 1: Distinguishing singular and plural forms	
	They sent us the missing parts. They sent us the missing part
Task 2: Distinguishing present tense from past tense	
	He stays near the river He stayed near the river
Task 3: Distinguishing the definite from the indefinite article	
	Why don't you read a newspaper? Why don't you read the newspaper?

Table 10.16 Combining sound, tense and grammar

Task 4 Listen and identify how many words are missing. Try to write down what you hear. Check that what you've written makes grammatical sense	
Sample questions	She took ____ river They ____ stream
Transcripts	She took a bath in the river She took the path to the river They rowed the boat up the stream They lowered the boat into the stream

Task 4 illustrates how listeners may use grammatical clues if they are not sure of the exact words they heard. Task 4 trained English teachers’ listening skills in distinguishing sounds of phonemes and develop grammatically correct sentences.

Grammatical, Syntactic and Semantic Skills

Research evidence shows that both native speakers and English language learners use a variety of clues to assist with comprehension. These clues include semantic clues (i.e. information provided by the context), syntactic clues (i.e. information provided by the grammatical structure of the sentences) and prosodic clues (i.e. information provided by the intonation and stress pattern of the sentences) (Berne, 2004). Thus, grammatical, syntactic and semantic skills are closely related to effective listening comprehension. Research studies show that native speakers and more advanced language learners rely more on semantic clues than syntactic clues and prosodic clues (Conrad, 1985; Ellis, Johnson, & Harley, 2000).

The LPATE training courses attempted to raise English teachers’ awareness of how to use grammatical (Task 5), syntactic (Task 6) and semantic clues (Task 7) to comprehend the listening passages by engaging teachers in doing a number of tasks. The tasks and sample test questions are listed in Table 10.17.

The message implied in the above three tasks is that listeners may make semantically correct inferences using semantic clues, when they are not sure of what they have heard. Task 5 shows that sentences should first be grammatically correct. In completing Task 6, participants were encouraged to make use of the subject–verb–(object) structure of the sentence. Thus, syntactic clues would be helpful in assisting

Table 10.17 Using grammatical, syntactic and semantic clues

Task 5 Which of the following is more likely to be the message?	
	He’s a guest in the house He’s a guested in the house She trussed the chicken and roasted it She trust the chicken and roasted it
Task 6 Using syntactic clues to connect sentences together	
	The short passage you are going to hear contains the following: a picture/a portable phone/the Titanic/a lifeboat/the lavatory/a night to remember/the Atlantic/trauma How might these words be linked together? After you listen, note down all the verbs that have been used
Task 7 Which sentence is more likely to be the message?	
	She rained for a long time She reigned for a long time He has carry-on luggage only He has carrion luggage only

the completion of listening tasks. Task 7 indicates that when sentences are grammatically and syntactically correct, attention should also be paid to the meaning of the sentences and see if the meaning makes sense.

In terms of cognitive abilities, all three tasks assess the cognitive abilities at the local level. In terms of linguistic skills and knowledge, these three tasks go beyond the conceptual meaning of words and phrases and pay attention to the meaning at the sentence level.

Micro- and Macro-Listening Skills

On the basis of the taxonomy of micro-skills in listening developed by Richards (1983), Brown (2007) developed a list of micro- and macro-skills in listening. The micro-skills pertain to skills at the sentence level, whereas macro-skills attend to skills at discourse level (Brown, 2007). Brown (2007) believes that being aware of such a taxonomy of skills would help English teachers to know what kind of skills the listeners need in order to acquire effective listening strategies. A list of micro- and macro-listening skills is presented below:

Micro-Skills

1. Retaining chunks of language of different lengths in short-term memory;
2. Discriminating among the distinctive sounds of English;
3. Recognising English stress patterns, words in stressed and unstressed positions, rhythmic structure, intonational contours, and their role in signaling information;
4. Recognising reduced forms of words;
5. Distinguishing word boundaries, recognise a core of words, and interpret word order patterns and their significance;
6. Processing speech containing pauses, errors, corrections, and other performance variables;
7. Processing speech at different rates of delivery;
8. Recognising grammatical word classes (nouns, verbs, etc.), systems (e.g., tense, agreement, pluralisation), patterns, rules, and elliptical forms;
9. Detecting sentence constituents and distinguish between major and minor constituents and
10. Recognising that a particular meaning may be expressed in different grammatical forms.

Macro-Skills

1. Recognising cohesive devices in spoken discourse;
2. Recognising the communicative functions of utterances, according to situations, participants, goals;
3. Inferring situations, participants, goals using real-world knowledge;
4. From events, ideas, etc., described, predict outcomes, inferring links and connections between events, deduce causes and effects, and detect such relations such as main idea, supporting idea, new information, given information, generalisation, and exemplification;
5. Distinguishing between literal and implied meanings;
6. Using facial, kinesic, body language, and other nonverbal cues to decipher meanings, and
7. Developing and using a battery of listening strategies, such as detecting key words, guessing the meaning of words from context, appealing for help, and signaling comprehension or lack thereof.

(Brown 2007, *Micro- and macro-skills of listening comprehension*, p. 308).

The above list illustrates how micro-skills involve the skills focusing on understanding the meaning of chunks of language as well as recognising the pronunciation of such chunks. Micro-skills focus more broadly on the meaning of an utterance in a context. The next section summarises the listening exercises practicing different listening skills on the LPATE courses.

Micro-Skill Listening Tasks

Table 10.18 lists a range of micro-skill listening tasks. These tasks addressed the following micro-skills in listening. More specifically, participants were expected to identify the missing words through relating the missing words to the respective pronouns (Task 8), relating the missing words to what had been mentioned (Task 9) and paying attention to verb tense (Task 13). Participants were also asked to identify larger stretches of sentence and put them in the right order (Task 11).

Generally, the above listening tasks attended to language at a micro-level by expecting participants to note down details in the listening passages.

Macro-Skill Listening Tasks

Table 10.19 presents three tasks used in the session on macro-listening skills. By doing Task 15 and Task 16, the participants had the opportunity to practise predicting what happened in the listening text and what speakers were probably talking about. Task 16, in particular, allowed participants to practise the abilities to detect links

Table 10.18 Micro-skill listening tasks

Task 8 Pronoun referents	
Sample questions	Sentence 1 “it”-
Sample transcripts	Sentence 1: I was listening to the Learning Curve programme last week and I found it very interesting because of my own situation
Task 9 Words or phrases which refer to something else that has been said	
Sample questions	Caroline Downs on the problems of getting appropriate education for an autistic child. Mike Collins, of the National Autistic Society, is with me. How typical is _____, Mike?
Sample transcript	Caroline Downs on the problems of getting appropriate education for an autistic child. Mike Collins, of the National Autistic Society, is with me. How typical is that problem, Mike?
Task 10 Identifying related words and ideas	
Sample questions	In the following extract you will hear an explanation of “highly structured approach” to learning in this school. Listen for four examples of “highly structured approach”
Task 11 Identifying cohesion	
Sample questions	The following clauses are utterances from the talk on Autism. Put them in the right order and write down the links that connect them to each other Autistic children often have a lot of energy At assembly, the children jog (The children) Go through what they call their posture exercise routine The school puts great emphasis on physical exercises
Task 12 Supplying missing words	
Sample questions	_____ went _____ initially and _____ expectations, _____. “Well, _____ expectations _____.”
Sample transcript	When I went to see the school initially and we were talking about expectations, they said “Well, we have expectations of the children.”
Task 13 Listening to the tape and identify the time markers and tenses of the verbs	
Sample questions	David _____ (identify) with children his own age group. He _____ (identify) with adults, or actually very young babies. _____ David _____ (start) this school, a most amazing thing _____ (happen)
Sample transcript	David has never identified with children his own age group. He identifies with adults, or actually very young babies. As soon as David started this school, a most amazing thing happened
Task 14 Cloze	
Sample questions	MC: I think it’s identified one of the cornerstones of good _____ for children with autism, and that is _____. We heard in the interview how _____ that is and certainly in _____ successful school setting for children with autism, structure is the cornerstone
Sample transcript	I think it’s identified one of the cornerstones of good practice for children with autism, and that is structure. We heard in the interview how important that is and certainly in any successful school setting for children with autism, structure is the cornerstone

Table 10.19 Macro-skill listening tasks

Task 15 Pre-listening discussion	
	<p>In small groups:</p> <ol style="list-style-type: none"> 1. Think of some reasons for listening to the video “learning about learning”: in what circumstance would you be listening to such a talk? 2. From your own life, think of one GOOD learning experience and one BAD learning experience. Describe these in your group. Have you any idea what made these experiences good or bad? 3. If you could ask several experts some questions about the learning process, what would you ask them? List your questions
Task 16 Matching the questions and answers in a listening text	
	<ol style="list-style-type: none"> 1. Participants are given eight answers given by an interviewee-David Gardiner and are asked to predict what questions were asked; then they listen to a listening text with questions only, and re-arrange these answers in the correct order, according to the questions asked 2. Then participants are given a list of interviewer’s questions and match them with the answers given by David Gardiner in the early part of the task
Task 17 Shadowing a spoken text	
	<ol style="list-style-type: none"> 1. Looking at the transcript, try to repeat after the speaker the exact words you heard. Continue until you feel comfortable with the speed of the recordings and can follow easily 2. Now put away the transcript and do the same thing by listening what the speaker says and repeating it: wait until the speaker has said two or three words, then start “shadowing”

between different stretches of sentences, by identifying themes, cohesive devices and organisational markers. Task 17 expected participants to comprehend the general meaning of the listening text and follow the text, by shadowing a piece of text.

Listening Tasks Incorporating Both Micro- and Macro-Listening Skills

In the training courses, there were also tasks integrating both micro- and macro-listening skills on topics familiar to English language teachers. These tasks included listening tasks on different topics, including academic talk (Task 18), understanding attitudes and opinions (Task 19), identifying different views in debate (Task 20), comparing schools and explaining differences to parents (Task 21) and listening task on educational issues (Task 22).

Task 18 focuses on taking notes and reflecting on how an academic talk is structured, to convey the meaning to the audience, as shown in Table 10.20. Three macro-listening skills are addressed in Task 18—‘making prediction before listening’, ‘detecting information from the listening text’ and ‘recognising how information is organised in the text’. Through taking notes of sub-headings as well as key issues under specific headings, participants could develop their capacities in identifying and collecting key information. Such skills are needed in situations such as attending lectures, professional development courses, or even school meetings. The post-listening

Table 10.20 Making predictions and detecting information from the text

Task 18 Listening to a presentation on the topic of ‘learning about learning’	
Pre-listening	<ol style="list-style-type: none"> 1. Think about the reasons for listening to such a text 2. Discuss good and bad teaching experiences
Taking extensive notes	<p>A. Listen to the first principle: new learning is shaped by learner’s prior knowledge Pay attention to how the presenters convince the listeners of the credibility of the theory. For example, discrediting other theories of learning? Or giving examples or evidence of their theory?</p> <p>B. Now we come on to the principle: “Learning is closely tied to particular situations” Think back to your purposes for listening. You may be attending a lecture to help you write an assignment, or you may be engaging in staff development. In any case, you will only get one chance to listen Write down some sub-headings and take notes</p> <p>C. Listen to the principle-“successful learning involves the use of numerous strategies” List below all the THINKING SKILLS you can come up with List below all the LISTENING SKILLS you can come up with:</p>
Post-listening	<ol style="list-style-type: none"> 1. Note down ways these principles could impact your classroom teaching 2. Note down ways these principles could impact YOUR classroom teaching if you implemented them 3. Note down any ways you think these principles apply to listening activities such as you have just done

tasks related the listening text to participants’ own teaching experience by asking them to reflect on how the strategies mentioned in the listening texts can be used in participants’ own teaching. This indicates that the training module is closely associated with English language teaching thus having the potential to improve proficiency of English language teachers.

Tasks 19 and 20 were two tasks on understanding attitudes and opinions. Task 19 below aimed to help participants understand attitudes and opinions using various sub-tasks, including summarising main arguments, evaluating attitudes on a five-point scale, note-taking, identifying supporting information and listening out for speaker mood. These sub-tasks tried to help participants practice their ability to summarise the main arguments and interpret the attitudes of the speakers. The micro-listening skills addressed were recognising communicative functions of the text, grasping the main ideas of utterance, inferring links between information (such as general and specific information). Micro-listening skills were also practised, for example, inferring the words showing the attitudes of speakers through stress and tone. In Task 20, micro-listening skills were practised through gap-filling exercises. Macro-listening skills were practised through noting down the required information (Table 10.21).

Table 10.21 Understanding attitudes and opinions

Task 19 Listening to an extended statement of opinion. Follow the argument and understand the attitudes of the speaker	
Summarising main arguments	Listen to the talk given on the radio by David Mellor on ‘the Olympics’. After listening once, get together in groups and try to reconstruct what the speaker’s main point(s) is/are The speaker believes that..... He thinks..... He argues that.....
Evaluating attitudes	Using the following charts to mark how you rate the speaker’s attitudes Agree disagree The speaker has strong opinions 5 ___ 4 ___ 3 ___ 2 ___ 1 The speaker is persuasive 5 ___ 4 ___ 3 ___ 2 ___ 1
Note-taking	Now listen to the talk about ‘the Olympics’ again. This time the recording you hear will have some pauses in it. You should take notes on what you hear in each segment, but there is no need to write down every word. Your aim will be to separate the main points from the supporting detail
Identifying supporting information	For rhetorical reasons the speaker makes several broad generalisations. Identify and note the examples, statistics and arguments which he uses to support the statements
Listening for the speaker’s mood	1. Listen and mark on the transcript below what you can hear: (a) Strong stress on particular words (b) High tone/pitch on certain words 2. Identify from the transcript examples of: (a) positive and negative adjectives (b) emotional words (c) colourful language (d) the speaker making the talk more personal Listen to the next segment and comment on it. Try to identify where the speaker express: (a) Surprise; (b) a sense of injustice; (c) a sense of struggle and hard work; and (d) disapproval
Task 20 Listening to debate in role	
Gap-filling exercises	You may remember that _____, we dutifully trumpeted _____. Its aim was to get more people _____ and show parents how to _____ in these skills. (Selected extract)
Note-taking	Make notes on this programme regarding the following information: How does Fantasy Football work? What is its aim? What is the advantage of it? How does it involve maths? How do the students at the school feel about it?
Information-gap listening exercises	Divide into two groups “Wolfes” and “Whites”. One group of you will listen to the points made by Professor Alison Wolfe, the other group to those made by Professor John White. Make notes and prepare the points as if you are Professor Wolfe or Professor White

Table 10.22 Drawing comparisons

Task 21 Listening to how parents choose secondary schools in English in order to make a comparison with the process of choosing schools in Hong Kong

Taking longer notes	Listen to the discussion and note as many points as you can understand in the following headings: Things to do before visiting a school Things to look for while visiting a school Other factors to consider
Making comparisons	Imagine a parent from England is coming back to Hong Kong and needs to know how to get their child into a secondary school here. Based on the information you have heard in the discussion, tell the parent in what way the system is similar here, and in what way it is different

Task 21 is closely related to English language teaching in Hong Kong (see Table 10.22). Participants were asked to compare different ways of selecting schools and presenting this to parents. Through drawing comparisons, participants were expected to extract main issues and identify key points of difference or similarity. To complete this task, participants needed to use macro-listening skills such as recognising the communicative function of a text, and inferring connection between events. After listening, participants were also expected to tell a parent the differences between the educational systems in Hong Kong and in England. Such an exercise made the task more authentic and participants had the chance to use what they had learned from the listening tasks and texts.

Task 22 in the listening module was on current topics in education (see Table 10.23). Micro-skills involved in Task 22 were recognising words through blank filling; and identifying grammatical mistakes in a text. Macro-skills in listening are grasping the main ideas, inferring speaker opinion and identifying connections between ideas. Such a task was educational related, and it provided an additional opportunity for participants to practise listening skills they used throughout the course.

Table 10.23 Current topics in education

Task 22 Current topics in education	
Blank filling	But first, the school and _____. Over the last couple of weeks, _____ has been a focus of attention in the education world. The RNIB published research showing that _____ and _____ youngsters are _____ at school and college (extract).
Understanding the speakers' experience	Richard tells of having attended 4 different schools. Listen to Richard's story and fill in some of the details of these. The schools he went to: His experience there/why he left: What do you think his experience taught him?
Understanding the speaker's opinion	Richard's opinion Where does Richard think money should be spent? Identify both general and specific information of Richard's opinion. How does he justify his opinion? How much time and money does he think needs to be spent?
Understanding education issues	What is the government policy/trend that Libby refers to?
Identifying the mistakes in a transcript	Listen and identify mistakes in the following transcript. But first- new kinds of school closer perhaps to business and industry than to the armed service ethos. In the league tables and top-scoring secondary school in English was the Thomas Telford School in Shropshire where many of its GCSE pupils got six or more ABC grades.
Post-listening discussion	What do you think 'inclusive' education should mean for teachers? What do you think 'inclusive' education should mean for disabled students? What do you think 'inclusive' education should mean for other students at the school? What are the pitfalls of inclusive education?

Summary and Conclusion

To sum up, the above tasks exposed English teachers to a range of listening task types and allowed them to practise their listening skills in comprehending the general meaning of a passage, as well as in catching the details of a passage, which address all four aspects of cognitive abilities as prescribed in the LPATE handbook (Government of the Hong Kong Special Administration, 2000), i.e. local processing, global processing, inferring and interpreting language in a larger context. To complete these tasks, participants needed linguistic skills and knowledge to derive conceptual meaning, propositional meaning, textual and rhetorical meaning and pragmatic meaning from the text.

The Reading and Listening modules were designed as a response to enhance participants' cognitive abilities and linguistic skills and competence that were prescribed in the LPATE Handbook. These two modules exposed participants to a large number of authentic tasks that were related to educational context and address reading skills and listening skills in a comprehensive way. The tasks provided the opportunity to enhance different aspects of cognitive abilities and linguistic skills and knowledge

(see Table 10.1). Thus, the two modules were intended to develop and enhance the Reading and Listening abilities that proficient English language teachers should have, in order to meet the Language Proficiency Requirement (LPR) as. These modules also had the potential to raise participants' awareness of the skills and strategies that are needed for second language learning. It was also the hope that the English teacher participants might apply these strategies in their own English language teaching.

The Chap. 11 describes and discusses the other three assessment areas in the LPATE training courses: Writing, Speaking and Classroom Language.

References

- Barton, L. (2002). *There's more to life... The guardian*. Retrieved 19 March from <https://www.theguardian.com/education/2002/mar/19/gcses2001.gcses>.
- Berne, J. E. (2004). Listening comprehension strategies: A review of the literature. *Foreign Language Annals*, 37(4), 521–531. <https://doi.org/10.1111/j.1944-9720.2004.tb02419.x>.
- Boers, F., & Lindstromberg, S. (2006). Cognitive linguistic applications in second or foreign language instruction: Rationale, proposals, and evaluation. In G. Kristiansen, M. Achard, R. Dirven, & F. J. R. D. M. Ibáñez (Eds.), *Cognitive linguistics: Current applications and future perspectives* (pp. 305–355). Berlin, New York: De Gruyter.
- Brown, H. D. (2007). *Teaching by principles: An interactive approach to language pedagogy*. White Plains, NY: Longman.
- Che, F. S. (2002). Teaching critical thinking skills in a Hong Kong secondary school. *Asia Pacific Education Review*, 3(1), 83–91. <https://doi.org/10.1007/bf03024923>.
- Conrad, L. (1985). Semantic versus syntactic cues in listening comprehension. *Studies in Second Language Acquisition*, 7(1), 59–69. <https://doi.org/10.1017/S0272263100005155>.
- Ellis, R. O. D., Johnson, K. E., & Harley, B. (2000). Listening strategies in ESL: Do age and L1 make a difference? *TESOL Quarterly*, 34(4), 769–777. <https://doi.org/10.2307/3587790>.
- Government of the Hong Kong Special Administrative Region. (2000). *Syllabus specifications for the language proficiency assessment for teachers (English language)*. Hong Kong: Hong Kong Special Administration Region Government.
- Grabe, W. (2008). *Models and more models of reading: Explaining reading in a second language: Moving from theory to practice* (pp. 83–106). Cambridge: Cambridge University Press.
- Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, 36(2), 193–202.
- Hedge, T. (2003). *Teaching and learning in the language classroom*. Oxford: Oxford University Press.
- Lund, R. J. (1991). A comparison of second language listening and reading comprehension. *The Modern Language Journal*, 75(2), 196–204. <https://doi.org/10.1111/j.1540-4781.1991.tb05350.x>.
- McDonough, S. H. (1995). *Strategy and skill in learning a foreign language*. London: Edward Arnold.
- Mecarty, F. H. (2000). Lexical and grammatical knowledge in reading and listening comprehension by foreign language learners of Spanish. *Applied Language Learning*, 11(2), 323–348.
- Mokhtari, K., & Thompson, H. B. (2006). How problems of reading fluency and comprehension are related to difficulties in syntactic awareness skills among fifth graders. *Reading Research and Instruction*, 46(1), 73–94. <https://doi.org/10.1080/19388070609558461>.
- Ozuru, Y., Dempsey, K., & McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction*, 19(3), 228–242.

- Park, G. P. (2004). Comparison of L2 listening and reading comprehension by university students learning English in Korea. *Foreign Language Annals*, 37(3), 448–458. <https://doi.org/10.1111/j.1944-9720.2004.tb02702.x>.
- Richards, J. C. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly*, 17(2), 219–240. <https://doi.org/10.2307/3586651>.
- Willis, J. (1981). *Teaching English through English: A course in classroom language and techniques*. Harlow: Longman.
- Yamashita, J. (2015). In search of the nature of extensive reading in L2: Cognitive, affective, and pedagogical perspectives. *Reading in a Foreign Language*, 27(1), 168–181.
- Zhang, J. L. (2001). Awareness in reading: EFL students' metacognitive knowledge of reading strategies in an acquisition-poor environment. *Language Awareness*, 10(4), 268–288. <https://doi.org/10.1080/09658410108667039>.

Barley Mak is Associate College Head of United College at The Chinese University of Hong Kong. She is a teacher educator, working with primary and secondary English language teachers at the undergraduate and postgraduate levels. Her publications have appeared in a considerable number of internationally referred journals. She was the founding Director of the Centre for Enhancing English Learning and Teaching (CEELT), has conducted various public-funded research projects, and has served on a number of prominent HKSAR teacher education committees.

Yangyu Xiao is a senior research assistant in the Department of Curriculum and Instruction at the Education University of Hong Kong. Her publication and research interests are in the fields of formative assessment, language curriculum and teacher education.

Chapter 11

The CUHK LPATE Training Courses: Writing, Speaking and Classroom Language



Barley Mak and Yangyu Xiao

Abstract This chapter focuses on the Writing, Speaking and Classroom Language Assessment modules—the three areas that are assessed by scales and descriptors in the LPATE. The scales and descriptors adopted in each LPATE paper, namely Writing, Speaking and Classroom Language Assessment, are first introduced, followed by a presentation of tasks that were used in different modules. This chapter focuses on how these tasks aided the development of the scales assessed in the LPATE, thus helping participants meet the LPR. From a wider perspective, this chapter describes how an enhanced grasp of the Writing, Speaking and Classroom Language Assessment modules may contribute to teacher professional development.

Introduction

The Writing, Speaking and Classroom Language Assessment papers in the LPATE assess candidates' production of written and oral language. The three tests are assessed by scales and descriptors which are central to competence in writing, speaking and teaching English through English in classrooms. It should be noted that although all sections of the Writing Module were criterion-referenced and assessed by means of scales and descriptors during the period described in this section, the LPATE revisions of 2006 amended the scoring patterns of the Writing Module in the changes that were made to the module and its assessment. Some forms of analytical marking were introduced in the revision process and promulgated once the revised version was implemented. These changes are described more fully by Urmston and by Drave in Section III.

B. Mak (✉)

United College, The Chinese University of Hong Kong, Sha Tin, Hong Kong
e-mail: barleymak@cuhk.edu.hk

Y. Xiao

Department of Curriculum and Instruction, Faculty of Education & Human Development,
The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, Hong Kong
e-mail: shirleyxiaoyy@gmail.com

As stated, the current section focuses on three modules—Writing, Speaking and Classroom Language Assessment. Tasks on the three modules are presented, and attention has been paid to show how these tasks enabled participants to enhance their language proficiency, as well as to satisfying the LPR as stated in the relevant scales of the Government’s LPATE handbook (Government of the Hong Kong Special Administration, 2000).

Assessing Writing, Speaking and Classroom Language—Scales, Descriptors and Levels

Developing scales and descriptors for language assessment is a development in language assessment, motivated by a need to produce more transparent and communicable test results than a numerical score (Hudson, 2005). Scales and descriptors are used in various criterion-referenced benchmark language tests, such as the Canadian Language Benchmark and the Common European Framework (Hudson, 2005). Scales and descriptors are helpful to both raters and language learners. Rating scales can facilitate raters to make evaluation decisions in a more reliable and manageable manner by providing raters with categorisations that raters can use (Lumley, 2002). Scales and descriptors are also helpful for curriculum design by offering a scaled summary of qualitative aspects of language use (Little, 2005).

Scales and descriptors vary according to aspects of language skills that are assessed (e.g. assessment of writing versus assessment of vocabulary). A brief summary of scales and descriptors of the three papers are presented in Table 11.1, with a detailed account of these scales and descriptors presented in Appendix A ‘The LPATE Writing Test—Assessment Scales Before 2007’. The scales and descriptors can be seen to relate to aspects particularly relevant to English language teachers’ language proficiency, rather than language proficiency in general.

Table 11.1 Overview of scales and descriptors in the LPATE

	Writing Test	Speaking Test	Classroom Language Assessment (CLA)
Scale 1	Grammatical accuracy	Pronunciation, intonation and stress	Grammatical accuracy
Scale 2	Organisation and coherence	Reading aloud with meaning	Pronunciation, intonation and stress
Scale 3	Task completion	Grammatical accuracy	The language of interaction
Scale 4	Ability to identify and correct errors	Organisation and cohesion	The language of instruction
Scale 5	Ability to explain errors	Interacting with peers	
Scale 6		Explaining language matters to peers	

There are similarities in the scales upon which performance in Writing, Speaking and Classroom Language Assessment was evaluated. All three papers pay attention to grammatical accuracy. Two papers relate to oral performance—Speaking and CLA include a *Pronunciation, Intonation and Stress* scale. The Speaking and Writing Tests include an *Organisation and Coherence* scale, as in both papers candidates are expected to organise their oral or written presentations logically and coherently.

On each assessment scale are there five levels (Level 1–Level 5). Level 5 is the highest level, which indicates that candidates have high language proficiency in the language assessed on a specific scale; Level 1 is the lowest level, indicating that candidates have little awareness of the respective language requirement, or demonstrate little capacity to meet the requirement. An example of the five levels for ‘*Grammatical Accuracy*’ in the Writing Test is presented in Appendix B ‘[The LPATE Speaking Test—Assessment Scales Before 2007](#)’.

The Writing Module

LPATE Writing Test and Assessment Scales

The Writing Test consists of two parts: Part I: *Expository Writing*; and Part II: *Error Correction and Explanation*. There were some modifications to the Writing Test when the LPATE paper was revisited and revised in 2007 (See Urmston, Chap. 14). Table 11.2 outlines the structures of the LPATE Writing Test before and after 2007.

As shown in Table 11.2, from 2007 onwards, three major changes were effected in the LPATE Writing Test:

Table 11.2 LPATE Writing Test before and after 2007

	The LPATE Writing Test before 2007	The LPATE Writing Test after 2007
Task types	<p>Part 1: Task 1 Expository writing Text: a text up to 200 words as a stimulus Task: Each candidate is given either a primary-focused task or secondary-focused tasks, depending on the teaching focus of the candidates at the time of application</p> <p>Part 2: Task 2A and 2B Correcting and explaining errors/problems Tasks: candidates are asked to correct 10–15 specified errors/problems in a student’s composition. They are then asked to explain a selection of these errors/problems</p>	<p>Part 1: Task 1 Composition Text: a text up to 200 words as a stimulus Task: Part 1 requires candidates to write one text of about 400 words (e.g. expository, narrative, descriptive, etc.) on a topic familiar to Hong Kong teachers (but not necessarily on education)</p> <p>Part 2: Task 2A and 2B Correcting and explaining errors/problems. Task 2A expects students to detect and correct errors in a student’s composition of appropriately 300 words Task 2B expects candidates to explain errors and problems in the format of gap-filling</p>

Table adapted from: Education Bureau of Hong Kong, 2007; Government of the Hong Kong Special Administrative Region, 2000

1. The topics for the writing task went beyond a purely teaching context.
2. Candidates were expected to have the capacity to write different text types apart from solely expository writing.
3. Instead of having to explain errors in a student text, candidates had to complete a separate task on error correction through gap filling.

The similarity lies in that both papers assess the candidates' capacity to produce a piece of writing and to identify and explain errors to students. The first aspect is closely related to candidates' own writing proficiency, whereas the second aspect relates to candidates' grammar-related pedagogical content knowledge. The structure of both LPATE papers echoes a recent study on teachers' perception of the LPATE—with the study indicating that qualified teachers need both language proficiency and pedagogical content knowledge to demonstrate that they are professional teachers (Coniam, Falvey, & Xiao, 2017).

As the LPATE training courses were provided between the years 2001 and 2005, the courses were designed to meet the LPATE requirement before 2007. As mentioned in the introduction, analytical marking was carried out for the new Task 2 from 2007. The scales and descriptors of the LPATE Writing Test before 2007 are summarised and presented in Appendix A '[The LPATE Writing Test—Assessment Scales Before 2007](#)'. Scales 1–3 were used to assess the 'expository writing' task, by focusing on *Organisation and Coherence*, *Grammatical Accuracy*, and *Task Completion*. Scales 4 and 5 were used to evaluate tasks on *correcting and explaining errors/problems*, by focusing on whether candidates were able to identify the errors and their capacity to explain the errors.

Awareness-Raising: Understanding the Writing Criteria

The Writing Module started with an awareness-raising exercise on understanding the criteria used to assess writing in the LPATE Writing Test. An understanding of the criteria was therefore intended to provide a prospective perspective so that participants would be able use the expected criteria, therefore being able to better shape their work (Sadler, 2005).

Participants taking the course were asked to study five expository writing samples. Such an approach—using writing exemplars—has been reported as a successful strategy in teacher professional development in that it helps teachers develop insights into teaching and assessing writing (Limbrick, Buchanan, Goodwin, and Schwarcz, 2010). Studying exemplars also enables English teachers to understand what writing at different levels looks like, as well as how to decide levels of different writing. Participants were asked to match five sample essays with five sets of grades, drawing on the scales and descriptors in the LPATE. Such an exercise was likely to raise participants' awareness of criteria and standards of the LPATE writing, before they approached specific writing tasks.

Simulation Writing Tasks

Simulation writing tasks was another approach used in the Writing Module. Simulation tasks create settings and tasks in such a way that they represent what are thought to be pertinent aspects of the real-life context (Shohamy, 1995). In second language education, simulation tasks are considered to generate rich authentic language, thus leading to active student engagement, and enabling students to use second language in the target culture (Oxford, 1997). Simulation tasks are also used in second language teacher education, as such tasks can be considered as authentic samples of pedagogical practice (Andrews, 2002). In the training module, the simulation tasks allowed English teacher participants to practise coping with situations that were likely to take place in real school settings, such as writing minutes, developing a discussion paper and writing an article for a newspaper, as Table 11.3 elaborates.

The three simulation writing tasks in Table 11.3 provided participants with authentic settings to write about. More specifically, Task 1 was about taking the minutes of a meeting. The simulation was that participants held a panel meeting to discuss students' work, as they indeed do at school, and took minutes of the panel meeting. Task 2 required participants to write a discussion paper to be submitted to the principal, on the basis of letters of complaint written by parents on writing instruction. Task 3 expected participants to write an opinion piece to a newspaper, in response to a criticism of the school's marking policy. Pre-task activities, including both lectures and discussion on the writing topic, were provided to facilitate the writing process. Such processes simulate the experience of writing that participants were likely to experience in school, thus making these writing tasks more authentic.

Participants' writing products were assessed according to Scales 1, 2 and 3 of the LPATE Writing Test, namely *Organisation and Coherence*, *Grammatical Accuracy and Task Completion*. It was intended that participants would become more aware of how to develop their writing to meet the expected standards as stated in the scales.

Error Correction

In the LPATE assessment, candidates were asked to correct errors in students' writing and explain errors to students. As error correction emerged as one of the weakest elements in the LPATE Writing Test from the perspective of candidate performance, the training module was developed to help participants develop strategies for correcting errors and provide them with practice in error correction.

Table 11.3 Simulation writing tasks

Task 1 Writing minutes	
Background	The principal of St. Luke's secondary school was concerned about the English language standards in the school after the school recently changed to CMI (Chinese Medium of Instruction). The school board decided that priority must be given to promoting English. At the end of the year, the principal decided to publish a newsletter to show how successful this year's theme has been. The newsletter, which will be sent to parents, included examples of student work. As there will not be space for a lot of student work to be shown, you have to decide which texts will be included
Pre-task	A lecture on 'how to write minutes of a meeting' Gap filling exercises on writing a minute
Simulation	Participants were given a set of "student" work, poems and stories, and instructions to hold a panel meeting to discuss the production of a school newsletter. Participants were divided into groups of 8 and given role cards (teacher A to E, secretary 1–2, and a chairman). They held a discussion and notes were taken
Writing task	The writing task is to write up the minutes of the meeting
Task 2 Developing a discussion paper	
Background	Your school has received two letters from parents with conflicting views on how student writing should be marked (whether teachers should mark every mistake or not). The principal has instructed the members of the English panel to re-examine its policy on marking and to prepare for a meeting to discuss the issues involved, at which a whole-school policy on marking will be adopted
Pre-task	Letters and extracts to be read at home
Simulation	There have been letters from parents complaining about the English department's marking policy. These letters were handed out, plus some other short extracts about error correction from teaching methodology textbooks
Writing task	Participants were asked to write a discussion paper to prepare for a meeting with the principal where the department's marking policy will be decided. A discussion paper should present the various options for error correction and marking, and should consider the pros and cons of these options. This paper should not propose one solution, but put forward ideas for discussion and final agreement by the department together with the principal
Task 3 Writing an article for a newspaper	
Background	A letter has appeared in the newspaper, accusing the school of laxity in its marking. You have been asked to write a response for publication in the paper. You should bear in mind that this will be read by the general public as well as teachers. It will be of particular interest to parents. You should write in a way that is appropriate to this audience. You should also present your points in such a way that people will be attracted by the subject, and their interest maintained
Pre-task	A lecture on developing a newspaper article and using reporting verbs Studying the format and style of newspaper articles in local educational supplements
Simulation	A letter has appeared in the newspaper, complaining about the school's marking policy. Participants were asked to write a response for publication in the local English language paper
Writing task	In this task, you have to direct yourself at a different audience, including the general public, teachers and parents of school-aged children. Pay attention to the style of a newspaper article; look at the ways it attracts and maintains readers' attention. Look at the types of language used: is it jargon known only to professionals, or more common terms?

Reviewing Grammatical Items

The training course on grammar started with a review of useful grammatical terms, to help participants review English grammar they would have probably learned at school or university. The grammatical items reviewed included nouns, pronouns, adjectives, verbs, verb tenses, prepositions, adverbs, determiners, sentence structures and clauses. The explicit revision on grammatical terms was intended to help participants to become familiarised with grammatical terms which could be used to explain grammar to students in the classroom. As English is not only the medium of instruction but also the objective of teaching, a good command of language knowledge is therefore necessary (Elder, 2001; Elder & Kim, 2014). Grammatical knowledge is a key aspect of teachers' content knowledge, as English teachers need to explain grammar to second language learners. In Section I of this volume (Coniam and Falvey, Chap. 1) concerning the initiatives for the introduction of the LPATE, it was noted how in Hong Kong in the 1990s a large proportion of English teachers were not subject-trained. In the context of the enhancement programme, it was therefore felt that an overview of grammatical items would provide all in-service English teacher participants with the chance to refine their grammatical knowledge.

Explaining Grammar to Students

The training module on grammar included a demonstration of how grammatical mistakes may be corrected, supported with exercises on error correction. Table 11.4 demonstrates language used to explain grammatical mistakes to students; four error correction and explanation tasks are presented in Table 11.5 as examples. Table 11.5 does not include all examples but rather serves as an example of how error explanation was practised in the training module.

Along with the explanations of grammatical items outlined above, there were grammatical exercises on correcting students' errors in writing and explain grammatical errors to students. Table 11.5 presents two such examples—on attributive clauses and on comparison, respectively. It was intended that these two examples would provide an insight into how error correction might be practised in the Writing Module.

The four example tasks on attributive clauses and on comparison give a flavour of the kinds of grammatical exercises participants conducted in the classroom. The tasks allowed participants to practise correcting errors, as well as to explain errors. To complete these tasks, participants needed to have relevant subject-matter knowledge as well as pedagogical content knowledge, so that they would be able to explain error correction to students in ways that were accessible and understandable. The exercises provide focused training on different aspects of English grammar, thus enhancing participants' capacity to identify and correct students' mistakes.

Table 11.4 Explaining grammatical mistakes to students

Grammatical items	Problem areas
Relative clauses	<i>Redundant relative pronoun</i> I agree that dating <i>which</i> is very time consuming <i>Missing relative pronouns leading to double sentences</i> Heroes and ordinary mortals are both human beings, they need to eat and sleep <i>That or which instead of who to refer to people</i> This is nothing for those students which are from a rich family, but what about the poor ones?
Pronouns	<i>Missing pronouns</i> I want to keep it as a pet, but I am afraid my mother won't allow ^ <i>Wrong case</i> Can you get he for me please?
Comparison	<i>Double comparison</i> I think a long and dull life is more preferable to a short exciting one <i>Missing comparison words</i> Today people in Hong Kong are ^ overweight and unfit than ever before
Negatives	<i>Correlative constructions (neither...nor.../not...either/none...neither)</i> She is not tall and not fat <i>Modals</i> Students are no need to bring mobile phones to school
Possessives	<i>Redundant possessive</i> Most school's don't hold dances because they do not want to encourage dating <i>Missing possessive</i> I am sure this will improve <i>Hong Kong</i> competitiveness <i>Time expressions</i> I have <i>seven days</i> holiday in December
Sentence structure	<i>Double sentences</i> The old woman did not say anything, she seemed every angry <i>Incomplete sentences</i> When a woman gets married. The woman must obey her husband <i>Faculty parallelism</i> I think a good teacher must have a good sense of humour, responsible and care
Word order	<i>Adjectives</i> She has <i>dark big</i> eyes <i>Compound subject</i> Last week I and my friends went camping on Cheung Chau

Summary

The Writing Module consisted of two parts—expository writing and error correction—broadly mirroring what was assessed in the LPATE Writing Test. The sessions on expository writing were intended to train participants to write in school settings through tasks simulating what would happen in schools. With regard to error correction, the training modules offered the chance to review grammatical items, as well as providing exercises for participants to correct and to practise explaining errors

Table 11.5 Sample error correction tasks

Task 4 Correcting the errors involving relative clauses	
Read the sentences from student compositions given below carefully. Some have errors involving <i>relative clauses</i> and some are correct. If the sentence is wrong, correct it. If it is correct, put a ✓ on the line	
In contrast, if someone who only stays at home, it is boring	(1) _____
She doesn't like girls they are more beautiful than she is	(2) _____
I saw a woman that was buying some fruit	(3) _____
Task 5 Explaining errors to students (attributive clause)	
Correct the errors in the following students' sentences, and discuss, in your own words, more fully the error problems 1. For some countries which too far from the equator, farming is not possible Analysis: _____ 2. But there are many parents that disapprove of their children dating Analysis: _____	
Task 6 Correcting errors involving comparison	
Read carefully the sentences from student compositions given below. Some have comparison errors and some are correct. If the sentence is wrong, correct it. If it is correct, put a ✓ on the line	
And I felt that beautiful morning like a wonderful dream	(1) _____
Singapore has a better infrastructure than Zhuhai	(2) _____
The creature looked liked very angry	(3) _____
The restaurant got so hot that some people became angrier	(4) _____
Task 7 Explaining errors to students. (comparison)	
Correct the errors in the following students' sentences, and discuss, in your own words, more fully the error problems 1. People in these poor countries don't have much money to spend as so Hong Kong people _____ 2. Students who get better grade in the examination should be allowed to go to university _____	

to their students. It will thus be appreciated that while the Writing Module targeted specifically the requirements as stated in the LPATE, it also attempted to impart and develop the knowledge needed by English language teachers.

The Speaking Module

LPATE Speaking Test and Assessment Scales

The Speaking Module was designed to fulfil the LPATE requirement before 2007 (a revision of the LPATE was conducted in 2007). A comparison of writing tasks in the LPATE Speaking Test before 2007 and after 2007 can be found in Table 11.6.

As shown in Table 11.6, from 2007, the reading of a poem was removed from the Speaking Test (see Urmston, Chap. 14 and Falvey & Coniam, Chap. 18, this volume). The CUHK Speaking Module, as provided in development courses between 2001 and 2005, still included a component on ‘reading aloud a poem’.

Three speaking topics were included in the training module ‘first day at school’, ‘pioneering journeys’ and ‘social and professional interaction’. Within each topic, a range of tasks targeted at enhancing participants’ speaking proficiency from different aspects was used. These tasks included reading a poem, reading a prose passage, phonology tasks, recounting a story, expressing view points, and speaking in groups. These tasks focused on speaking proficiency in line with the six scales specified in Table 11.1 and Appendix B ‘The LPATE Speaking Test—Assessment Scales Before 2007’).

Speaking Tasks in the Training Module

Altogether 36 speaking tasks were used in the training module. A summary of these 36 tasks on three topics are listed in Table 11.7 (first day at school), Table 11.8 (pioneering journeys) and Table 11.9 (social and professional interaction). In each table, the purposes of the tasks and scales upon which these tasks were evaluated are presented. The tables show that these tasks covered all six scales, namely *Pronunciation, Reading Aloud with Meaning, Grammatical Accuracy, Organisation and Coherence, Interacting with Peers*, and *Explaining Language Matters to Peers*, as stated in Table 11.1. These scales shared similarities with the three levels of oral proficiency identified by Iwashita, Brown, McNamara, and O’Hagan (2008) in assessing oral language proficiency in the pilot TOFLE iBT, namely linguistic

Table 11.6 The LPATE Speaking Test—before and after 2007

	The LPATE Speaking Test before 2007	The LPATE Speaking Test after 2007
Task types	Part 1: Task 1A Reading aloud a poem Task 1B Reading aloud a prose passage Task 1C Telling a story/recounting an experience/presenting arguments Part 2: Task 2 Group interaction	Part 1: Task 1A Reading aloud a prose passage Task 1B Recounting a personal experience or presenting arguments based on a stimulus Part 2: Task 2 Group discussion

resource (grammatical accuracy, grammatical complexity and vocabulary), phonology (pronunciation, intonation and rhythm), and fluency (pause, repair, speech rate and number of syllabus produced in the utterance). An additional aspect in the LPATE

Table 11.7 Speaking tasks and their aims—first day at school

Tasks	Purposes of tasks	Scales addressed
Poem: 'First day at school' by Roger McGough		
Task 1 Reading a children's poem to explore meanings, word-play and feelings	To help participants to read a poem with understanding and expression	Scale 2
Task 2 Practising elements of stress and intonation	To help participants to read the poem aloud so that it can be clearly understood	Scale 1
Task 3 Identifying short and long vowels	To raise awareness of the basic spelling rules	Scale 1
Task 4 Making a recording of a reading of 'First day at school' by Roger McGough	To allow participants to demonstrate that they can read aloud a selected text with meaning and in such a manner that it would arouse the interest of the listeners	Scale 2
Prose 'My first day of school' Ladies' home journal Sept. 1998		
Task 5 Talking about an episode in the past	To help participants develop oral fluency in recounting an experience	Scale 4
Task 6 Reading aloud examples of authentic spoken English in a group, focusing on attitudes and feelings expressed by the speaker	To help participants become familiar with the rhythms and idioms of spoken discourse, and give practice for scale 2—reading aloud with meaning	Scale 2
Task 7 Studying examples of authentic English and abstracting pattern of usage	To build language awareness, in particular related to tense used in talking about the past, the difference between spoken and written styles, and transitional words and phrases	Scale 3
Task 8 Transforming a text from present tense to past tense	To practise accurate grammatical use	Scale 3
Task 9 Recording a short informal speech remembering your first day as an English teacher	To demonstrate the ability to recount an episode from the teachers' own experience in a coherent manner, using appropriate tenses and transitions	Scale 4
Phonology IPA and phonetic transcription Voiced and voiceless consonants		
Task 10 Reading a phonetic transcription in IPA (International Phonetic Alphabet)	To assess participants' knowledge of the International Phonetic Alphabet	Scale 1
Task 11 Matching spellings with articulation charts	To identify the phonemes of English from their conventional spelling	Scale 1
Task 12 Deriving a phonological rule for the information of words with -s and -ed endings	To show that phonological processes are regular and can often be captured by simple principles	Scale 1

Speaking Module focused on how participants communicate and interact with peers. This aspect was specifically relevant to the LPATE Speaking Test.

The first 12 tasks were intended to strengthen participants' oral competence, as assessed by Scales 1–4 in the LPATE. These tasks were built on school-related themes and drew attention to the following skills: reading aloud with understanding, reading with accurate pronunciation and grammatical use and recounting one's own experiences.

Tasks relating to 'social and professional interaction' are presented in Table 11.9. These tasks were more advanced in term of the complexity of the language required; they were also more closely related to participants' daily encounters in teaching.

It will be noted that Scales 5 and 6 were the two major scales that were practised in Unit 3, where participants had the opportunity to practise '*Interacting with Peers*' (Scale 5) and '*Explaining Language Matters to Peers*' (Scale 6). Similar to tasks used in the Writing Module, simulation tasks and role-plays were used so that participants could practise spoken English in real lifelike situations.

Six speaking tasks are presented in Table 11.10 as an illustration of how different speaking tasks were used to address each of the six assessment scales. Whereas the samples do not represent all the tasks that were used in the training modules, they give readers a flavour of how different scales were practised and addressed in the Speaking Module.

As Table 11.10 shows, Task 12 focused on *Pronunciation, Stress and Intonation*; the task expected participants to transcribe words into IPA (the International Phonetic Alphabet) as their tutor read out different words. Thus, participants needed to have both a good knowledge of IPA as well as an understanding of how sounds are related to pronunciation.

Task 1 was a task on '*Reading Aloud with Meaning*'. After reading a poem, participants needed to answer a series of questions which would help them relate meaning to the way they read the poem. Participants were also asked to think about where to pause—in order to make the meaning clearer to the audience.

Task 8 expected participants to practise reading out a passage in a grammatically accurate way. Although the task only required participants to transform present tense into past tense, the task was intended to remind participants about the need to pay attention to tense while speaking.

Task 21 asked participants to plan a talk before making the talk, thus drawing attention to how a talk should be coherent and properly organised.

Task 27 drew out explicitly patterns of conversations in different school situations, as well as ways to tackle difficulties in conversation. This task provided a chance for participants to practise communicating with peers, as well as raising participants' awareness of potential obstacles in conversations and how they might be dealt with.

Task 32 was a group discussion task where participants were expected to explain their ideas to their peers.

The above six examples illustrate how tasks with different focuses were used in the LPATE training module to enhance participants' competence in speaking. More importantly, these tasks drew explicit attention to the key issues in effective speaking—thus raising participants' awareness about ways to improve their oral English.

Table 11.8 Speaking tasks and their aims—pioneering journal

Tasks	Purposes of tasks	Scales addressed
Talk: Reminiscences of migrating to Australia		
Task 13 Listening to the reminiscences of a migrant to Australia in the 1950s and take notes	To identify the stage of talk	Scale 4
Task 14 Analysing the transition in the text	To identify cohesive devices	Scale 4
Poem: In the desert Prose: Overseas Chinese		
Task 15 Analysing a poem's syntactic structure in groups	To observe the different relationship between parts of complex and compound sentences	Scale 3, 5, 6
Task 16 Marking the pauses and run-on in a poem and suggesting ways to read it aloud	To read aloud with meaning	Scale 1, 2
Task 17 Reading aloud "where did the overseas Chinese come from"	To demonstrate ability to read aloud with meaning, with clear pronunciation. To apply what has been learnt about sentence structure to use correct phrasing in reading aloud	Scale 1, 2
Using meta-language		
Task 18 Constructing a parallel text in the group	To reinforce grammar and meta-language, and practise discussing language matters with peers	Scale 3, 6
Task 19 Completing a blank cloze in groups	To demonstrate understanding of English syntax by reconstructing a text	Scale 3, 6
Constructing narrative		
Task 20 Constructing a simple narrative based on a sentence	To use relative clauses and embedded structures to embellish a story	Scale 3
Task 21 Giving a short talk about a migrant journey	To practise a talk from notes structuring the narrative round stages and key words	Scale 3, 4
Phonology The vowels of English: Pure vowels and diphthongs		
Task 22 Distinguishing and producing 3 vowel phonemes in English	To learn to distinguish three similar phonemes in English (/e/, /ei/ and /æ/). These three phonemes create considerable difficulties for Cantonese speakers	Scale 1
Task 23 Practising vowel production	To make a clear distinction between words with similar vowel sounds	Scale 1
Task 24 Transcribing into IPA one verse of the poem "in the desert"	To recapitulate and consolidate what has already been learned about English vowels. Provide a starting point for new material on vowel length in English	Scale 1

Table 11.9 Speaking tasks and their aims—social and professional interaction

Tasks	Purposes of tasks	Scales addressed
<i>Group dynamics</i>		
Task 25 Warming up. Free form discussion, asking for advice	To focus attention on some of the features of natural conversation and discussion	Scale 5
Task 26 Conventions of conversation	To understand the different rules of conversation in Cantonese and in English	Scale 5
Task 27 Social behaviour in conversations	To understand the social and interpersonal factors affecting communication in conversations	Scale 5
Task 28 Roles in discussions	To understand the roles people should play in discussion	Scale 5
Task 29 Role-play card game	To practise performing certain essential functions in discussion To identify language used to perform these functions	Scale 5
Task 30 Observing turn-taking in authentic speech	To understand how native speakers take turns to talk, and compare it with the average dialogue in a text book	Scale 5
<i>Simulation</i>		
<i>Choosing books as class readers</i>		
Task 31 Reading aloud from selected texts	To practise reading aloud with meaning.	Scale 1, 2
Task 32 Justifying your choice of texts used as textbooks for primary and secondary school students	To practise expressing one's views and justifying the explanation	Scale 5, 6
Task 33 Reaching an agreement in the group	To practise negotiation and reaching an agreement	Scale 5, 6
<i>Phonology</i>		
<i>Difficulties for Cantonese speakers</i>		
Task 34 Pronouncing initial single consonants	To practise consonants difficult for Cantonese speakers	Scale 1
Task 35 Practising the/s/ and /esh/sounds in a variety of phonological environments	To distinguish the two similar sounds	Scale 1
Task 36 Voiced and voiceless final stop	To understand and distinguish the voiced and voiceless final stop	Scale 1
Task 37 Distinguishing l and r	To increase participants' awareness of the articulation of these two sounds	
Task 38 Reciting a short poem	To practise reading aloud with meaning	Scale 1

Table 11.10 Sample speaking tasks

<i>Scale 1: Pronunciation, stress and intonation</i>	
Task 12	Deriving a phonological rule for the formation of words with –s and –ed endings
	Listen to your teacher’s pronunciation of the following words. Transcribe them into IPA, paying careful attention to the sounds used in the added inflections Hops, rises, bathed, pages, called, cats, carted, laughs, roamed, waifs, waves, helped, watches, balls, poked, sings, laughed, loves, faced, kisses, wished, paths, watched, hobs, loved, cads, homes, praised, needs, paged, pays, rained, seeks, tags, banged, hears, fines, cared, wishes, paid, raided, keeps, catches, docks, begged, dogs, stabbed, bathes, hates, robs, beiges
<i>Scale 2: Reading aloud with meaning</i>	
Task 1	Reading a children’s poem to explore meanings, word-play and feelings
	Poem: First day at school (by Roger McGough) A millionbillionwillion miles from home Waiting for the bell to go (To go where?) Why are they so big, other children? So noisy? So much at home they Must have been born in uniform Lived all their lives in playgrounds Spent the year inventing games They don’t let me in. Games. They are rough, that swallow you up
	Answer the following questions: 1. Who is the speaker in this poem? What things tell us that this is a child speaking? What experience is he/she describing? 2. Listen to the tutor reads the poem aloud and mark where there are pauses and where lines run on. How helpful is the punctuation? 3. Explain the relationship between lines 5, 6 and 7. 4. Which things do you think give the child a feeling of comfort? 5. Try reading the poem aloud in your group, thinking about how you want it to sound
<i>Scale 3: Grammatical accuracy</i>	
Task 8	Transforming a text from present tense to past tense
	Read the following passage and transform it into past tense I live in a small village which has only one school—it’s more than a mile outside the village, in the country side. My friends and I usually walk to school—it can be very wet and cold in winter, but there’s no other way to get there. We love it when it snows a lot and school’s closed. There is one bus but it doesn’t leave the village until 9 a.m., which is when school starts, so if we took that we’d be late. The walk to school is uphill all the way. It’s hard work when the wind is in your face
<i>Scale 4: Organisation and cohesion</i>	
Task 21	Give a short talk about a migrant journey
	Think up the basic outline of a story you want to tell about a journey. It may be a story of a family member or someone you have heard of, or your own story if you have had such an experience. Write the outline down the page the same way as you wrote your sentence. Add in some other parts you want to tell as notes and asides (This task is assessed according to Scales 3 and 4.)
<i>Scale 5: Interacting with peers</i>	
Task 27	Social behaviour in conversations
	Free discussion: How is it decided who will speak in these different types of verbal interaction? e.g., staff meetings; department meetings; committee meetings; groups of colleagues having lunch together What social and interpersonal factors lie behind any differences? Suggest ways of dealing with the following conversational difficulties Someone who always dominates the conversation A situation where you find yourself the only person talking A situation where you have an idea but everyone else is talking so much that you can’t get a word in
<i>Scale 6: Explaining language matters to peers</i>	
Task 32	Justify your choice of texts used as textbooks for primary and secondary school students
	Group discussion Decide as a group which text you will choose to be the class reader. You should consider all the extracts, and be prepared to present your decision to the other groups, explaining why you have chosen this particular text and saying why you would not use the others. Quote from the texts to back up your own view. (This task is assessed according to scales 5 and 6.)

Summary

The Speaking Module in the LPATE training courses provided a series of tasks to address the six scales assessed in the LPATE Speaking Test. These tasks were intended to support participants in improving their language proficiency through providing the opportunity for practice and through raising their awareness of the diverse set of elements associated with what can be broadly termed ‘effective speaking’.

The Classroom Language Module

LPATE CLA Paper and Assessment Scales

The Classroom Language Assessment (CLA) paper expects teachers to demonstrate their ability to communicate to students with appropriate grammar and with appropriate pronunciation, stress and intonation. The CLA paper consists of the assessment of two live lessons on two separate school days. The assessment is conducted by two assessors, with each assessor observing a single class teaching period. The CLA intends to assess language ability rather than teaching methodology. It should be noted that after the 2007 revision of the LPATE, the compulsory assessment of two lessons ceased, and was replaced by one assessment plus a number of randomly selected assessments (see Drave, Chap. 14; Falvey & Coniam, Chap. 18).

Language ability in the CLA is assessed on four scales, namely (1) grammatical accuracy, (2) pronunciation, stress and intonation, (3) the language of interaction, (4) the language of instruction. A description of these scales and descriptors can be found in Appendix C ‘[The LPATE CLA Paper—Assessment Scales Before 2007](#)’. *The Language of Interaction* and *The Language of Instruction* are the two main aspects of language ability specifically related to language teaching. Grammatical accuracy and pronunciation are embedded in the assessment of both *The Language of Instruction* and *The Language of Interaction*.

An Overview of Classroom Language Tasks

Table 11.11 outlines CLA tasks and on which scales these tasks were assessed. These tasks offered participants the opportunity to practise their classroom language: the purpose of each task is outlined in Table 11.11 alongside the nature of different tasks. Participants worked in small groups on each task, with group work rather than micro-teaching being used in the classroom language module, as it was felt that regular group work was more likely to offer participants a greater opportunity to practise different aspects of classroom language. As the course participants were all

Table 11.11 Classroom language tasks and their purposes

Tasks	Purposes of tasks	Scales addressed
Task 1 Understanding the scales and descriptors for CLA	To raise awareness of the different purposes of the language we use in the classroom	Scale 3 and 4
Task 2 Understanding the scales and descriptor for CLA	To appreciate the differences between Form and Function in the CLA scales	Scale 1 and 2
Task 3 Analysing the stages in a lesson	To identify the language signalling different stages of lessons	Scale 4
Task 4 Practising introducing stages of the lesson	To practise the language of introducing different stages of lessons	Scale 3 and 4
Task 5 How we communicate with our students (pause, stress and intonation)	To understand the effects of pause, stress and intonation on the language teachers use to give instruction to students	Scale 2 and 4
Task 6 How we communicate with our students (hand gestures)	To understand the gestures teachers use to communicate with students	Scale 4
Task 7 How we communicate with our students (stress, rhythm and gestures)	To understand how language and gestures can be used together to communicate with students	Scale 2 and 4
Task 8 How we communicate with our students (positive and negative language)	To practise ways of using positive and negative language to discipline students	Scale 4
Task 9 Using various language to manage and discipline class	To practise ways of using different language to manage and discipline students	Scale 4
Task 10 Discussion about the teaching of grammar	This task will lead participants to think about their own attitude towards the teaching and presentation of grammar	Scale 1 and 4
Task 11 Analysing a segment of a video recording of a pre-service teachers	This task will help participants think about different approaches to the presentation of teaching points	Scale 4
Task 12 Commenting on the different types of presentations	To understand the effects of different ways of presenting	Scale 4
Task 13 Observation of recorded elicitation techniques	This task will lead participants English teachers to think about what they are doing when they elicit language	Scale 3
Task 14 Assessing various method of elicitation	To identify and evaluate elicitation techniques	Scale 3
Task 15 Promoting oral interaction with and among students	To help participants English teachers think about activities that can promote oral interaction with students	Scale 3
Task 16 Question types	To practise different question types and think about the different responses that might be elicited by different question types	Scale 3

(continued)

Table 11.11 (continued)

Tasks	Purposes of tasks	Scales addressed
Task 17 Responding to students' answers	To identify the responding strategy that can elicit correct answers	Scale 3
Task 18 Jigsaw reading session: sharing opinions on error-correction	To encourage reflection on the decision-making processes involved in correcting errors	Scale 3
Task 19 Planning patterns of interaction in the classroom	To help participants English teachers think about the language needed for different patterns of interaction	Scale 3

in-service teachers, they were also asked to reflect on their classroom language when they taught in their own classes.

The different purposes of the tasks above demonstrate that The Language of Interaction (Scale 3) and The Language of Instruction (Scale 4) were two major areas of focus in the CLA module. The practice of the first two scales, namely Grammatical Accuracy (Scale 1) and Pronunciation, Stress and Intonation (Scale 2) were embedded in tasks on the language of interaction and the language of instruction. The following parts illustrate the tasks which were used to develop the above four scales, using selected classroom language tasks.

Understanding the Classroom Language Assessment Criteria

Similar to the Writing Module, the CLA module started with a session which aimed at enhancing participants' understanding of the scales and descriptors of the CLA paper. Participants had the chance to study classroom samples, which they could relate to the scales and descriptors developed for the LPATE; they also reflected on their own understanding of the role of classroom language in English language teaching, as shown in Table 11.12.

In Task 1, participants were given a range of classroom activities and asked to categorise them either as *'The Language of Interaction'* or as *'The Language of Instruction'*. Task 1 was intended to enable participants to reflect on their own classroom experience and thus become more aware of the two functions (i.e., the language of interaction and the language of instruction) of their classroom language. In Task 2, participants worked on a series of classroom situations and evaluated them in terms of on which scales the responses from teachers might be problematic. With the assistance of specific examples, Task 2 could therefore be seen to help with understanding of which scale a teacher's classroom language might be evaluated on.

Table 11.12 Sample task on understanding the CLA scales and descriptors

Task 1	Distinguishing between ‘The Language of Interaction’ and ‘The Language of Instruction’
	<p>Look at the extract from Syllabus Specifications for the Language Proficiency Assessment. Categorise these following activities as either ‘the language of interaction’ or ‘the language of instruction’, according to your understanding of the CLA</p> <p><i>Asking the students to be quiet</i> <i>Announcing the purpose of today’s lesson</i> <i>Telling the students to take out a different book</i> <i>Introducing a new grammar structure</i> <i>Giving the answers to an exercise completed earlier</i> <i>Answering a student’s question about what you taught yesterday</i> <i>Correcting a student when s/he gives a wrong answer</i> <i>Asking questions to find out if students have understood the lesson</i> <i>Commenting on a student presentation</i> <i>Encouraging a student to try to answer</i></p>
Task 2	Distinguishing Form and Function in the CLA scales
	<p>Scale 1 (Grammatical accuracy) and Scale 2 (Pronunciation, Stress and Intonation) are designed to assess the FORM of what the teacher says, whereas Scale 3 and 4 are designed to assess how well and appropriately the teacher’s language fulfils its FUNCTION. Read the scales and discuss the following examples of teacher utterances. Are these good examples? Identify the utterances that have problems, and say which scales would be referred to in each case</p> <p>The teacher asks, “who has bring the story book today?” The students all put up their hands to show they have</p> <p>The teacher repeated says the word “children” as [tʃɪdɪn]</p> <p>The teacher says, “If I added the past participle to this phrase, what difference would it make to its meaning?” to a low level F2 class</p> <p>A shy student has given an answer. It is slightly wrong. The teacher wants her to try again and get the right answer. S/he says, “No. Wrong. Do it again, now”</p> <p>A student says, “I forget bring my book.” The teacher says, “You forgot to bring your book? Well, can you share with your partner?”</p>

The Language of Instruction

Three key aspects assessed in the Language of Instruction include: *signalling*, *giving instructions* and *presenting*. The material used to develop and build on these elements of language will now be briefly outlined.

Signalling

Signalling is a classroom strategy which involves indicating the stage of a lesson by using appropriate language signals. Before completing Task 3—analysing the stages in a lesson—a range of techniques on signalling the different stages of class

Table 11.13 Tasks on signalling the stages in a classroom

Task 3	Analysing the stages in a lesson
	Watch the highlights of an English class and identify the stages in the lesson. Make a brief note to describe each stage
Task 4	Practising introducing stages of a lesson
	In groups of 4 or 5, get one copy of the jumbled lesson plan. You need to: Step 1: Re-order the jumbled lesson plan and Step 2: Compare your lesson with another group (member), then think of the next step in the lesson When you have completed this, appoint one person in the group as the “teacher”, and practise how you would lead your students from one stage of the lesson to the next. You should try to ensure they understand your instructions, and, if you think it is necessary, that they can see the relationship between parts

from Willis’ (1981) *Teaching English through English* were introduced to the participating English teachers. These included a wide set of language examples related to classroom organisation (such as greeting, starting or ending a lesson, checking attendance, instructions on using equipment). These examples provided participants with resources that they would consider using in their own classrooms as ways of signalling the changes in the stages of a lesson.

Participants worked on two tasks related to signalling: Task 3—analysing the stages in a lesson and Task 4—practising introducing stages of a lesson, as shown in Table 11.13.

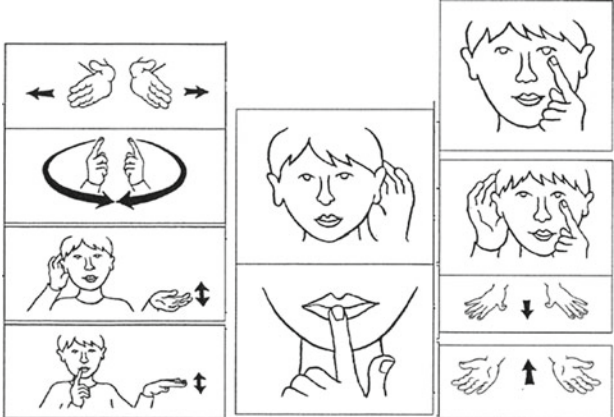
Task 3 aimed at raising participants’ awareness of how teachers might use different language signals at different stages of lessons. Task 4 offered the authentic experience of practising introducing a lesson.

Giving Instructions

English teachers need to give instructions when conducting activities, giving homework, and managing classrooms. Effective instruction communicates meaning and also creates impact; the language of instruction also needs to be modified so that it suits the ability level of students. Tasks 5–9 demonstrate how different aspects of language support giving instruction in the classroom (see Table 11.14).

Task 5 focused on intonation, stress and pause. Intonation, stress and pause can impact on communication as they bear a communicative load, act as grammatical cohesion, as well as having a pragmatic function (Pickering, 2001). A study comparing the native-speaking teaching assistant and ITAs’ (International Teaching Assistants) intonation in class settings show that ITAs’ presentations and the ways they talk to students indicated a limited number of negotiations with students (Pickering, 2001). Task 5 aimed at helping teachers become aware how pauses, stress and intonation may help with communicating meaning to students.

Table 11.14 Instruction giving tasks

Task 5	Intonation, stress and pausing
	English teachers first complete a pre-task by listening to three classroom extracts and marking down: (1) where the teacher pauses; (2) stressed words; and (3) intonation patterns
Task 6	Hand gesture: Match the following gestures with their meanings
	<p>Sit down, please; Stand up, please; Listen; Quite, please; Watch and listen; Watch; Everyone; Say it quietly, please; Say it louder, please; Say the whole sentence, please</p> 
	Adapted from: Garden, B and Gardner, F. (2000) <i>Oxford Basic Classroom English</i> , Oxford: Oxford

Task 6 focused on hand gestures. In classrooms, learners interpret teachers' gestures in conjunction with verbal language input in order to learn successfully (Sime, 2006). In the training module, participants were given a series of pictures and potential meanings of gestures and were asked to match the pictures with the accompanying meanings. Task 6 thus drew explicit attention to how hand gestures may be used to express meanings.

After developing an understanding of how classroom language may be used in different ways to produce different impacts, participants were also asked to work on a series of commands in the classroom, and to express similar meanings in a different way. Participants listened to a recording of a series of commands and commented on whether the words indicated a mild or a strong command. Such a practice enabled teachers to extend their classroom language and think of different ways of talking to students.

For primary and junior secondary students, it was suggested that simple phrases and expressions might be more appropriate—so that students might understand them more easily. As for advanced students, simple commands can be built upon with a view to providing more variety and to familiarise students with different types of language. Hence, one focus of the CLA course was centred on helping teachers to practise extending simple language so that it would cater to the needs of students of different levels.

Table 11.15 Presenting tasks

Task 11	Analysing a segment of a video recording of a pre-service teacher
	<p>1. Read the benchmarking criteria for Scale 4 of the CLA component (language of instruction)</p> <p>2. Watch the segments of language presentation conducted by pre-service teachers in their teaching practice. Make notes on the content of the lesson: what is happening in each stage? Describe the approach to presenting teaching points</p> <p>As you watch the teachers presenting a new language point in their classrooms, note down the stages they go through. The following notes will help you</p>
Task 12	Commenting on different types of instructions
	<p>Comment on the different types of presentations used by the teachers. Consider the clarity of the explanations, the coherence of the presentation, the level of understanding required by the students, the degree of involvement of the students, and the meaningfulness of the explanations. Also consider the planning required by the teacher, and the unknown factors involved when presenting new materials</p>

Presenting

Presenting involves organising spoken language so that information is presented to learners in a coherent and accessible way, e.g. explaining a grammar point, a vocabulary point or a concept. Two sample tasks are provided in Table 11.15.

The Language of Interaction

Effective interaction in the second language classroom is likely to facilitate student learning, thus contributing to learning from two possible angles: first, students complete the tasks successfully, with the assistance of teachers; and second, students achieve a level of independence, due to the learning experience received in the first aspect (Gibbons, 2003).

There is a wide variety of studies on the language of interaction—for example, on the language of teachers, teacher and learner beliefs, factors which shape the interaction, social and cultural background, and psychological aspects (Tsui, 2001). With regards to the language used by teachers, issues needing to be considered include teacher questions, learner responses, teacher feedback, and turn-taking behaviour (Tsui, 2001). In the LPATE courses, the language of interaction was assessed from three aspects: *eliciting*, *responding* and *giving feedback*.

Table 11.16 Sample eliciting task

Task 14	Assessing various methods of elicitation
	<p>Watch the video and evaluate each elicitation technique:</p> <p>(a) What is the teacher trying to elicit from the students in each extract</p> <p>(b) How successful is the teacher on each occasion</p> <p>(c) Which technique works best and why?</p> <p>(d) Which lessons require the students to give the most complex responses? Which the least complex?</p> <p>(e) Which eliciting techniques brought out the most meaningful responses? Which the least?</p> <p>(f) Which brought out the most structured responses?</p>

Eliciting

Eliciting strategies include asking questions, modifying or reformulating a question, providing clues and hints in order to help students provide an appropriate response, as well as encouraging students to ask questions and respond to each other. Table 11.16 demonstrates an *eliciting* task.

The above task was intended to raise teachers’ awareness of the effects of eliciting techniques and to what extent these techniques were helpful, so that English teachers might consider how these techniques might be used in the classroom.

Table 11.17 Sample tasks on promoting interaction among students

Task 15	Promoting oral interaction with and among students
Task description	<p>1. Consider the following questions:</p> <p>(a) How much do your students speak in class?</p> <p>(b) How do you prepare speaking activities in class?</p> <p>(c) What particular problems do your students have with speaking in class?</p> <p>In what ways do you encourage real communication to take place in your lessons?</p> <p>2. Below is a list of activity types designed to help learners develop oral fluency. Think of an example of each type from your own experience</p> <p>(1) Information gap; (2) ranking; (3) jigsaw; (4) guessing; (5) problem-solving; (6) role-play; (7) group discussion; (8) task-based activity; and (9) prepared speech</p>
Task 16	Question types
Task description	<p>1. Display and reference questions</p> <p>Participants read two sets of questions-display and reference questions along with the responses they elicit. Then they are asked to think about the following questions:</p> <p>What are the advantages and disadvantages of using “display” and “reference” questions?</p> <p>What kinds of questions do you use in classroom?</p> <p>2. Participants studied different types of questions and the answers that the questions elicited</p> <p>3. Participants were given a set of answers and were asked to write questions to elicit these answers</p>

In addition to eliciting a response from an individual student, English teachers also need to promote interaction among students. Task 15 is a general task on promoting responses among students whereas Task 16 focuses specifically on questioning, as Table 11.17 illustrates.

Questioning is an important strategy in the classroom that elicits responses from students. Tsui (2001) suggests that teachers should modify their questions if they fail at first to elicit responses, paraphrase difficult words and simplify the syntax if the questions are too complex. Tsui (ibid) believes that teachers should study samples of effective and ineffective questions by watching videos of their own lessons. In this vein, Task 16 illustrates a task on using questioning as an eliciting strategy.

Responding and Giving Feedback

In the classroom, teachers need to respond to students in various ways: seeking clarification, giving confirmation, and asking for repetition. They are also expected to provide feedback and comment on students' responses. Table 11.18 demonstrates two tasks outlining responding to students and giving feedback.

Table 11.18 Responding and giving feedback tasks

Task 17	Responding to students' answers
	<p>Read the following samples of student-teacher interaction. Identify teachers' aims in each segment:</p> <p>1. T: Have you got any ambitions? What is your ambition? S: Nurse T: You want to be a nurse. (to another student) Yours? Yes, you. Yes, have you got any ambitions? (students laugh) Nothing? (to another student) You S: A teacher T: To be a teacher. O.K</p> <p>2. T: What is the reason? S: Because he can play tennis and also ping-pong, also drive the sports and mm he can speak German T: She can speak</p>
Task 18	Jigsaw reading session: sharing opinion on error correction
	<p>1. Participants read three texts on three ways of correcting students' errors in spoken language. Then they discuss how students' errors in spoken English should be corrected</p> <p>2. After your reading, reflect different ways of correcting errors. Work in groups and complete the sentences below:</p> <p>(a) Teachers should interrupt learners who have made a mistake or error when...</p> <p>(b) Teachers might wait until later to give feedback when...</p> <p>(c) Some errors should be left uncorrected by the teacher, for example, ...</p> <p>(d) Teachers can help learners to self-correct by...</p>

Task 17 was an awareness-raising exercise. Task 17 allowed participants to read examples of teachers eliciting language from students with the objective of becoming aware of how different classroom language elicits different responses from students.

Task 18 provided the opportunity for participants to discuss ways of providing feedback on students' errors and for them to reflect on strategies concerning correcting students' errors in different situations.

To sum up, the tasks in the CLA module focused mainly on Scale 3 (*The Language of Interaction*) and Scale 4 (*The Language of Instruction*). These tasks had the potential to raise participants' awareness about using their own language for different purposes in the classroom through exposing them to different examples and chances for practice.

Summary and Conclusion

The Writing, Speaking and Classroom Language Assessment modules were designed to help participants fulfil the LPR of the LPATE by focusing on the respective assessment scales laid out in the LPATE handbook (Government of the Hong Kong Special Administration, 2000). The tasks provided in the three modules were closely associated with using language in classrooms and provided participants with various opportunities to practise the language used in the school and classroom contexts. These tasks also raised participants' awareness of their written and spoken language in the context of teaching, thus contributing to the development of their language proficiency in the school context.

We now move on to Section III.

Appendix A: The LPATE Writing Test—Assessment Scales Before 2007

Scales	Descriptions
Scale 1 for Task 1	Organisation and coherence (Aspects assessed in this scale include: the development of ideas; the extent to which propositions are justified and elaborated or illustrated with examples to enhance meaning; the extent to which the text is coherent; the extent to which the text displays full audience awareness and appropriate register)
Scale 2 for Task 1	Grammatical accuracy (Aspects assessed in this scale include: the extent to which grammatical structures are accurate; the extent to which a wide range of structures are used)

(continued)

(continued)

Scales	Descriptions
Scale 3 for Task 1	Task completion (Aspects assessed in this scale include: the extent to which the content demanded of the writer by the task is presented; the extent to which the task is fulfilled; the extent to which the writers display sensitivity to the audience)
Scale 4 for Task 2A	Correcting errors/problems in a student's composition (Aspects assessed in this scale include: the language ability to identify and correct errors; the ability to deal with complex discourse-level errors; the percentage of errors that have been corrected)
Scale 5 for Task 2B	Explaining errors/problems to students (Aspects assessed in this scale include: the language ability to explain errors; the ability to explain complex discourse-level errors; the percentage of errors that have been fully and appropriately explained)

Table adapted from *Syllabus specifications for the language proficiency assessment for teachers—English language*, 2000

Appendix B: The LPATE Speaking Test—Assessment Scales Before 2007

Scales	Descriptions
Scale 1 for Task 1A and 1B	Pronunciation, Stress and Intonation
Scale 2 for Task 1A and 1B	Reading aloud with meaning (which includes: speed of delivery and pausing; sensitivity to the text and to the audience; and the use of paralinguistic features to communicate the text)
Scale 3 for Task 1C	Grammatical accuracy (which includes: accuracy in grammatical structures and the range of structures)
Scale 4 for Task 1C	Organisation and cohesion (which includes: the use of means for connecting utterances; how relationship among concepts and ideas are expressed, signalled, and whether there is confusion; flow of ideas in discourse; the range of vocabulary used)
Scale 5 for Task 2	Interacting with peers (which includes: the ability to talk easily, confidently and knowledgeably with peers in a professional manner; control over the conversational strategies of initiation, turn-taking, responding and disagreeing; ability to keep discussion focused)

(continued)

(continued)

Scales	Descriptions
Scale 6 for Task 2	Explaining language matters to peers (which includes: the ability to organise discourse or explain a students' language problems to peers; control over and familiarity with a wide range of appropriate meta-language without confusing peers; ability in producing appropriate examples to illustrate explanations; whether explanations are coherent and easy to follow)

Table adapted from *Syllabus specifications for the language proficiency assessment for teachers—English language, 2000*

Appendix C: The LPATE CLA Paper—Assessment Scales Before 2007

Scales	Descriptions
Scale 1	Grammatical accuracy (which includes: the accuracy of grammatical structure; the occurrence of inaccurate expressions)
Scale 2	Pronunciation, stress and intonation (which includes: the accuracy of pronunciation; whether and to what extent there are first language characteristics; sentence stress and intonation patterns; and the effectiveness of communication)
Scale 3	The language of interaction (which includes: the level of linguistic awareness and sensitivity to student responses; the ability to react in an appropriate linguistic manner to students' initiation; the language ability to be aware of and to react to students' responses even if these are incomplete or lacking in coherence; and whether and to extent teachers have language problems that impede communication)
Scale 4	The language of instruction (which includes: the ability of using English as the language of presentation; the ability to organise discourse and use appropriate signalling devices in order to alert students to the various stages of a presentation; and whether and to what extent classroom instructions are clear, comprehensible, and appropriate for the level of the class)

Table adapted from *Syllabus specifications for the language proficiency assessment for teachers—English language, 2000*

Appendix D: Levels and Descriptors for Scale 2—Grammatical Accuracy in Writing

5	Grammatical structures are always accurate, with no occurrence whatsoever of non-idiomatic or other inappropriate expressions. There is access to a wide range of structures, which can be invoked at any time. Any ‘mistakes’ that occur can be categorised as lapses rather than systematic errors
4	Grammatical structures are mostly or always accurate. In isolated instances, non-idiomatic or otherwise inappropriate expression may occur but communication is never impeded
3	Grammatical structures are greatly accurate but errors may occasionally occur when more complex structures are attempted. Comprehension is seldom impeded. Some complex structures are attempted
2	Grammatical errors occur regularly and may sometimes impede the readers’ understanding. Few complex structures are attempted
1	Most of the texts contain grammatical errors, causing comprehension to break down completely at times. Access to basic structures is clearly adequate and communication with reader is often impeded

Table adapted from *Syllabus specifications for the language proficiency assessment for teachers—English language*, 2000, p. 53

References

- Andrews, S. J. (2002). Teacher language awareness and language standards. *Journal of Asian Pacific Communication*, 12(1), 39–62. <https://doi.org/10.1075/japc.12.1.04and>.
- Coniam, D., Falvey, P., & Xiao, Y. (2017). An investigation of the impact on Hong Kong’s English language teaching profession of the language proficiency assessment for teachers of English (LPATE). *RELC Journal*, 48(1), 115–133. <https://doi.org/10.1177/0033688216687455>.
- Elder, C. (2001). Assessing the language proficiency of teachers: Are there any border controls? *Language Testing*, 18(2), 149–170. <https://doi.org/10.1177/026553220101800203>.
- Elder, C., & Kim, S. (2014). Assessing teachers’ language proficiency. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1–17).
- Garden, B., & Gardner, F. (2000). *Oxford basic classroom English*. Oxford: Oxford.
- Gibbons, P. (2003). Mediating language learning: Teacher interactions with ESL students in a content-based classroom. *TESOL Quarterly*, 37(2), 247–273. <https://doi.org/10.2307/3588504>.
- Government of the Hong Kong Special Administrative Region. (2000). *Syllabus specifications for the language proficiency assessment for teachers (English language)*. Hong Kong: Hong Kong Special Administration Region Government.
- Hudson, T. (2005). Trends in assessment scales and criterion-referenced language assessment. *Annual Review of Applied Linguistics*, 25, 205–227.
- Iwashita, N., Brown, A., McNamara, T., & O’Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, 29(1), 24–49. <https://doi.org/10.1093/applin/amm017>.
- Limbrick, L., Buchanan, P., Goodwin, M., & Schwarcz, H. (2010). Doing things differently: The outcomes of teachers researching their own practice in teaching writing. *Canadian Journal of Education/Revue Canadienne de l’éducation*, 33(4), 897–924.

- Little, D. (2005). The common european framework and the European language portfolio: Involving learners and their judgements in the assessment process. *Language Testing*, 22(3), 321–336. <https://doi.org/10.1191/0265532205lt311oa>.
- Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, 19(3), 246–276. <https://doi.org/10.1191/0265532202lt230oa>.
- Oxford, R. L. (1997). Cooperative learning, collaborative learning, and interaction: Three communicative strands in the language classroom. *The Modern Language Journal*, 81(4), 443–456. <https://doi.org/10.1111/j.1540-4781.1997.tb05510.x>.
- Pickering, L. (2001). The role of tone choice in improving ITA communication in the classroom. *TESOL Quarterly*, 35(2), 233–255. <https://doi.org/10.2307/3587647>.
- Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education*, 30(2), 175–194. <https://doi.org/10.1080/0260293042000264262>.
- Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics*, 15, 188–211. <https://doi.org/10.1017/S0267190500002683>.
- Sime, D. (2006). What do learners make of teachers' gestures in the language classroom? *IRAL—International Review of Applied Linguistics in Language Teaching*, 44, 211–230.
- Tsui, A. B. M. (2001). Classroom interaction. In R. Carter & D. Nunan (Eds.), *The Cambridge guide to teaching English to speakers of other languages* (pp. 120–125). Cambridge: Cambridge University Press.
- Willis, J. (1981). *Teaching English through English: A course in classroom language and techniques*. Harlow: Longman.

Barley Mak is Associate College Head of United College at The Chinese University of Hong Kong. She is a teacher educator, working with primary and secondary English language teachers at the undergraduate and postgraduate levels. Her publications have appeared in a considerable number of internationally referred journals. She was the founding Director of the Centre for Enhancing English Learning and Teaching (CEELT), has conducted various public-funded research projects, and has served on a number of prominent HKSAR teacher education committees.

Yangyu Xiao is a senior research assistant in the Department of Curriculum and Instruction at the Education University of Hong Kong. Her publication and research interests are in the fields of formative assessment, language curriculum and teacher education.

Part III

The LPATE: A High-Stakes Assessment in Operation (2001–2007)

Alan Urmston and Neil Drave

This section, consisting of four chapters, looks at the development and administration of the LPATE from the perspective of two HKEAA officers, one former and one current.

The LPATE was launched as a public examination in March 2001 for serving teachers of English in Hong Kong primary and secondary schools who had to attain the Language Proficiency Requirement (LPR) before September 2005. The first two chapters—Chaps. 12 and 13 by Urmston—report on the operation of the LPATE during the crucial years of 2001 to 2005, and the review/revision project that was carried out once the deadline for the attainment of the LPR by serving teachers had passed. The third—Chap. 14 by Drave—extends the detail reported in Chaps. 12 and 13 with a specific focus on standards setting at the HKEAA, while Chap. 15, also by Drave, examines perceptions of the LPATE in the media.

Chapter 12

The Operation of the LPATE (2001–2005)



Alan Urmston

Abstract This chapter looks at the operationalisation of the LPATE from its launch in 2001 through to 2005, when the decision was made to revise the Assessment. After an initially slow start, where approximately 400 candidates took the Assessment, perhaps because teachers embraced the possibility that the Education Department (as it was then) would not enforce the LPR, the LPATE went from strength to strength, with the candidature increasing steadily to over 2000 each administration, resulting in the HKEAA administering the Assessment twice per year (March and September) from 2003 to 2005. The chapter describes the technical aspects of test design and the operational complexities of running the Assessment in the midst of clear opposition to it from stakeholders. Issues to be discussed include the sociological and educational impact and consequences of such a high-stakes assessment.

Introduction

Amidst concerns over falling standards of language proficiency of students in Hong Kong, which emerged during the late 1980s and early 1990s, recommendations were made by the Hong Kong Education Commission in its Report No. 6 (Hong Kong Education Commission, 1995) that:

... minimum language proficiency standards should be specified for all new teachers to ensure that they can teach competently through the chosen medium of instruction; and that all new Chinese or English language teachers should, as from a certain cut-off date, be required (a) to have a high level of academic attainment in Chinese or English, and (b) to have completed satisfactorily professional training in the teaching of Chinese or English as a subject. (p. viii)

A. Urmston (✉)
Faculty of Humanities, English Language Centre,
The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
e-mail: alan.urmston@polyu.edu.hk

© Springer Nature Singapore Pte Ltd. 2018
D. Coniam and P. Falvey (eds.), *High-Stakes Testing*,
https://doi.org/10.1007/978-981-10-6358-9_12

239

In this regard, the Commission passed to the Government's Advisory Committee on Teacher Education and Qualifications (ACTEQ) two proposals for action to address the issue of the language competency of serving teachers:

1. Levels of language and professional competence ('benchmark' qualifications) to be established for all language teachers [Note 1];
2. Minimum language proficiency standards to be met by all teachers in their chosen medium of instruction.

(Falvey & Coniam, 1997, p. 17.)

Following these recommendations, in April 1996, two teams of consultants were engaged to develop language benchmarks for teachers of English and teachers who teach through the medium of English, and for teachers of Chinese and Putonghua and teachers who teach through the medium of Chinese [Note 2], respectively (Falvey & Coniam, 1997). This chapter will look briefly at the development of the LPATE (as this has been discussed previously in Coniam and Falvey (2013) and Section IV of this volume) and then go on to describe the implementation of the Assessment from the perspective of a Subject Officer of the HKEAA responsible for it between September 2001 and July 2005.

Timeline of Development of the LPATE (English Language)

Figure 12.1 presents a brief chronological account of the key stages in the development of the Language Proficiency Assessment for Teachers (English Language) or LPATE Government of the Hong Kong Special Administrative Region (2000).

As a result of the initial study looking at the requirements for a benchmark test of English language for teachers (Coniam & Falvey, 1996), recommendations were made that there should be a five-point scale of proficiency, to be known as the Benchmark Levels or later the LPATE Levels, and that the middle point on the scale, i.e. Level 3, should be the actual benchmark or minimum standard required for a teacher to be able to teach English in primary and secondary schools. There was some discussion as to whether there should be a higher standard required for secondary teachers given, on average, the higher demands on the proficiency level at which teachers would need to teach at secondary level. In the end, it was decided that adopting a different standard would prevent teachers teaching in primary schools who had attained the required level for primary teaching (but not secondary) from moving into the secondary sector, unless they retook the LPATE or otherwise attained the required level for secondary. For more discussion on the setting of the Benchmark Levels, see Section I (this volume).

April 1996	ACTEQ decides that the language benchmarking project should initially deal with the language proficiency of lower-secondary school teachers of English Language. Consultants embark upon study to determine requirements of benchmark test of teachers' language proficiency.
July 1996	Publication of consultants' report (Coniam and Falvey, 1996) recommending that the benchmark test should comprise the following components: (1) Formal tests of Reading, Writing, Listening, Speaking (2) Direct assessment of classroom language
September 1996	Consultants' report accepted by ACTEQ. English Language Benchmark Subject Committee set up under the auspices of the Hong Kong Examinations Authority (HKEA) (renamed Hong Kong Examinations and Assessment Authority (HKEAA) in July 2002) to finalise specifications, task types and associated descriptors for pilot exercise, known as Pilot Benchmark Assessment (English) (PBAE).
January 1998	Moderation Committees set up by HKEA to develop tests of Reading, Writing, Listening and Speaking for PBAE.
November 1998 – January 1999	Hong Kong Education Department carries out Classroom Language Assessment component of PBAE.
February 1999	Pen-and-paper tests of PBAE carried out by HKEA.
September 1999	Submission of PBAE report and recommendations to ACTEQ setting out criteria for testing instruments, scales and associated descriptors for testing instruments and establishing benchmark levels.
February 2000 – March 2000	After further pilot tests of primary and upper secondary English teachers, consultants recommend that one set of scales and tests be used for all primary and secondary teachers of English Language. The recommendation subsequently endorsed by ACTEQ and accepted by Government.
June 2000	Government announces that teachers who have received appropriate training in the teaching of English will be exempted from taking the LPATE.
November 2000	Publication of Syllabus Specifications for the Language Proficiency Assessment for Teachers (English Language) (Government of Hong Kong Special Administrative Region, 2000).
March 2001	First administration of the LPAT
June 2001	Release of results of LPAT 2001
September 2001	This author took over as Subject Officer of LPATE
March 2002	Second administration of LPAT

Fig. 12.1 LPATE Timeline

The First Administration of the LPATE

The first administration of the LPATE was carried out in March 2001; the results are shown in Table 12.1.

The first noticeable aspect of the results shown in Table 12.1 is the relatively low number of candidates, considering that approximately 13,000 English language teachers in Hong Kong schools would need to attain the Language Proficiency Requirement (LPR) before the end of 2005–2006 school year, either through the LPATE, through Government-approved training courses or by exemption due to

Table 12.1 Results of LPATE 2001

Paper	Number of candidates	Number achieving level 3 or above	Percentage achieving level 3 or above (%)
Reading	398	341	86
Writing	387	129	33
Listening	376	257	68
Speaking	351	178	51
Classroom Language	93	83	89

Table 12.2 LPATE candidature (2002–2005)

Administration	Candidature
2002 (March)	708
2003 (March)	1968
2003 (September)	2739
2004 (March)	2177
2004 (September)	1494
2005 (March)	1115
2005 (September)	1445
2006 (March)	953

already being subject-trained. Thus, the LPATE had not proved to be very popular with English language teachers. There were several possible reasons for this, one of which was that teachers were adopting a ‘wait and see’ approach to the initiative in the hope that the HKSAR Government would backtrack and abandon the initiative. There were other reasons, of course, and these will be discussed later in this chapter. As it turned out, though, the candidature increased dramatically over the next four years, as shown in Table 12.2.

In 2003, the Hong Kong Education Department instructed the HKEAA to provide two administrations of the LPATE per academic year from 2003–2004 through 2005–2006—the deadline for teachers to attain the LPR (Fig. 12.1).

Test Development

Test development at the HKEAA (or HKEA as it then was) follows a standard approach in which each *paper* or test (for the LPATE there were, and still are, four) is developed separately by a team consisting of a Chief Examiner, a Setter and two Moderators. The Subject Officer as was (now called Assessment Manager—a HKEAA employee) serves as secretary to this team, or *moderation committee* as it is known. In September 2001, the moderation committees for the LPATE (for the Reading, Listening, Writing and Speaking Tests) were mostly already in place, having worked on the 2001 tests. The members of the moderation committees for the LPATE were drawn from Hong Kong tertiary institutions, particularly those which provided English language teacher education programmes and were subject to approval from the LPATE Subject Committee [Note 3]. The process of test or paper development is quite standard and can be represented by the flow chart in Fig. 12.2 (see also Choi & Lee, 2009).

The above process, assuming a test administration in March, would normally begin in May or June of the previous year. This author took over the role of Subject Officer in September 2001, meaning there was not sufficient time to go through the whole process of test development to have a full set of tests ready by March 2002.

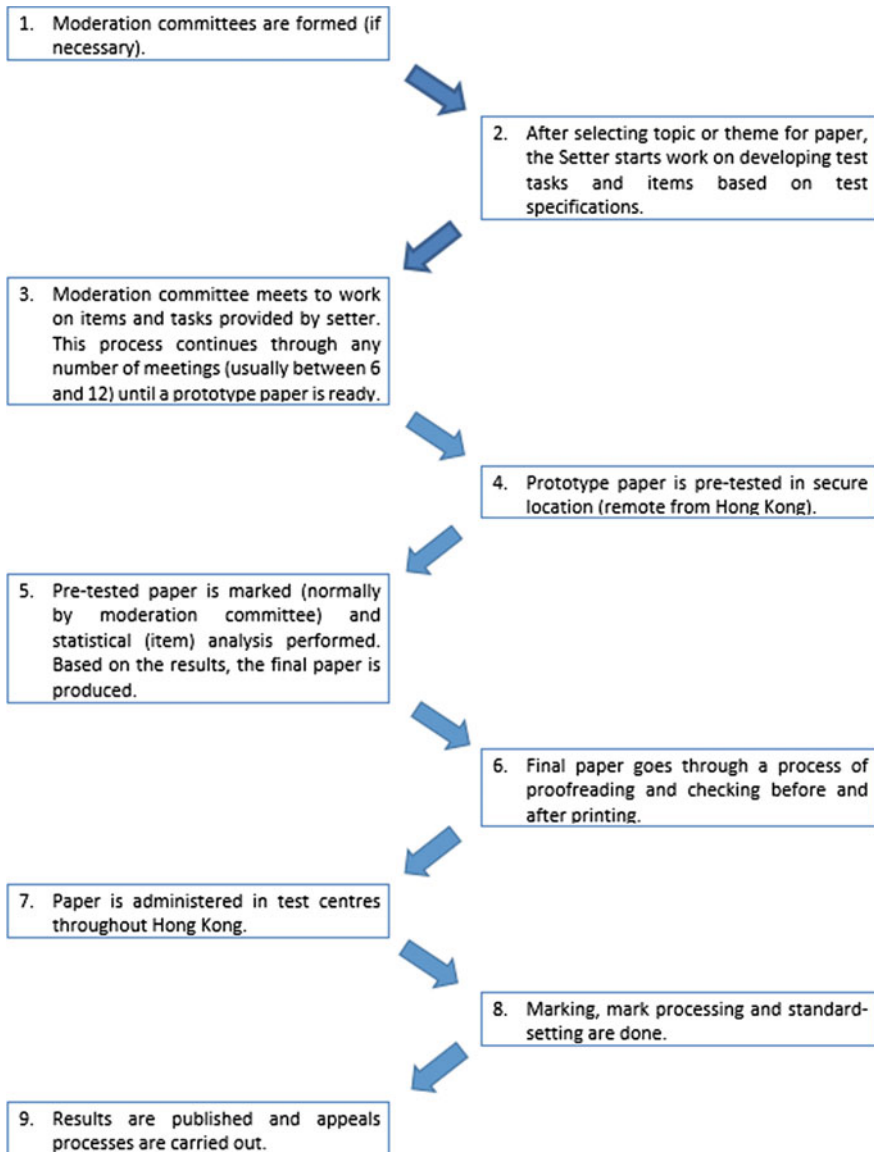


Fig. 12.2 HKEAA process of test paper development

Therefore, it was fortunate that in the process of development of the March 2001 tests, a reserve set had been prepared. This set then became the March 2002 set and the moderation committees worked on ‘polishing’ the tests to ensure that they would be ready to be administered in March 2002. After the administration of the LPATE in March 2002, including the post-administration tasks of marking, standard-setting,

Table 12.3 Pre- and post-testing of LPATE (2002–2005)

Date	Pre- or post-test	For LPATE administration
November 2001	Pretest	March 2002
November 2002	Pretest	March 2003
November 2003	Post-test	Sept 2003
December 2003	Pretest	March 2004
April 2004	Post-test	March 2004
June 2004	Pretest	Sept 2004
December 2004	Pretest	March 2005
June 2005	Pretest	Sept 2005

result reporting and appeals marking, the preparation process for the March 2003 administration began. Having gone through the full process of test development, administration, marking, and reporting in 2002–2003, it was announced by the Education Department that from the start of the 2003–2004 academic year, there would be two administrations of the LPATE per year, in September and March, until the deadline by which all serving teachers should have met the Language Proficiency Requirement, i.e. the start of the 2006–2007 academic year. The reasoning behind this move was that teachers needed to be provided with more opportunities to take the Assessment given the anticipated rush to attain the LPR by the deadline. This essentially doubled the workload of the LPATE team and made producing high-quality tests that much more difficult. The moderation process relies on the participation of part-timers, mainly teacher educators and university teachers, so most of the test development work was done in the early summer, when teaching had finished and before moderation team members went away on summer holidays. Then, prototype tests were made ready for pre-testing in November or December. With the increase to two administrations per year, the schedule had to be revised and operations squeezed. The schedule of pre-testing during this period, shown in Table 12.3, is indicative of this.

Pre-testing of the LPATE was done for two main purposes:

- For trialling of tests and test items to ensure that the prototype tests of Reading, Writing and Listening were of the required level of difficulty and that test items and tasks met specifications of standards in terms of reliability.
- For standard-setting: the prototype tests of Reading and Listening were administered to a cohort of test takers who also did ‘anchor tests’ of these skills for which the standards had previously been set. A process of test equating was then carried out using Many-Facet Rasch Analysis (MFRA) to provide information on potential cut scores for the Expert Judgement Panel, which would meet to set the test standards after the live tests had been administered [Note 4].

There were two occasions on which post-testing rather than pre-testing was done (see Table 12.3):

November 2003—it was not possible to carry out pre-testing in June 2003 (for the Sept 2003 tests) due to the outbreak in Hong Kong and elsewhere in East Asia of Severe Acute Respiratory Syndrome (SARS), so instead, the papers underwent a more extensive process of trialling in Hong Kong and for standard-setting, post-tests were carried out on the Reading and Listening Tests in November 2003.

April 2004—pre-testing in December 2003 (for the March 2004 papers) involved significantly fewer test-takers than the numbers expected, so while there proved to be adequate numbers to provide trials of the test papers, to gather sufficient data for standard-setting, post-tests were carried out on the Reading and Listening Tests in April 2004.

LPATE Test Design

The design structure of the LPATE as used from 2001 until 2006 is shown in Tables 12.4, 12.5, 12.6, 12.7 and 12.8.

The Reading Test, as shown in Table 12.4, consisted of two parts, a multiple-choice cloze and a reading comprehension section. The setting, marking and standard-setting of this paper were relatively straightforward and typically around two-thirds of test takers achieved a Level 3 or above on the Test.

Candidates found the Writing Test to be the most challenging, with typically between 30 and 40% of candidates achieving an overall Level 3 or above. The main difficulty for most candidates was in Part 2, in which they had to correct a selection

Table 12.4 Design of LPATE reading test

Test specification	Assessed skills/abilities/scales	Proficiency requirement determination method
Part 1: Multiple-choice cloze (30 min) One text of up to 750 words or two texts of no more than 800 words. Approx. 25 items	Cognitive abilities: <ul style="list-style-type: none"> • local processing • global processing • inferencing • interpreting language in a larger context 	Test papers were marked analytically and candidates awarded a score. Cut scores for the five proficiency levels were determined by a combination of Rasch equating [Note 5] and expert judgement [Note 6]. Each candidate was then assigned a proficiency level of 1–5 according to these cut scores.
Part 2: reading comprehension (60 min) One text of 1500–2000 words or two texts of 750–1000 words each on topics relevant to English language teaching and education. Approx. 35 items in 20–30 questions of various types including open-ended short questions, table or diagram completion tasks and multiple-choice items	Linguistic skills and knowledge: <ul style="list-style-type: none"> • conceptual meaning • prepositional meaning • textual or rhetorical meaning • pragmatic meaning 	

Table 12.5 Design of LPATE writing test

Test specification	Assessed skills/abilities/scales	Proficiency requirement determination method
Part 1: task 1 expository writing (approx. 60 min) A text, relevant to teaching, of up to 200 words was used as stimulus for a writing task of approximately 300 words	Part 1 <ul style="list-style-type: none"> • Organisation and coherence • Grammatical accuracy • Task completion 	Test papers were marked independently by two markers who assigned proficiency levels (1–5) according to descriptors on the five scales indicated. Candidates were assigned the means of these levels
Part 2: tasks 2A and 2B correcting and explaining errors/problems in a student's composition (approx. 60 min) Candidates were asked to correct 10–15 specified errors/problems in a student's composition. They were then asked to explain a selection of these errors/problems	Part 2 <ul style="list-style-type: none"> • Detection of errors at the morphosyntactic and discourse levels • Explanation of errors at the morphosyntactic and Discourse Levels 	

Table 12.6 Design of LPATE listening test

Test specification	Assessed skills/abilities/scales	Proficiency requirement determination method
Listening and responding to one or more segments of spoken discourse of approximately 30 min on a topic relevant to English teachers. Total time approx. 60 min with 5 min preparation and 10 min writing-up time Question types included open-ended short questions, table or diagram completion tasks, multiple-choice items, post-listening written responses at or above sentence level	As for the reading test	As for the reading test

of the errors in a student's composition and then explain some of those errors. In fact, a breakdown of the results shows that candidates performed relatively well on the correction task (approximate pass rates of 70%), but relatively poorly on the explanation task (approximate pass rates of 35%). Although it is difficult to draw conclusions as to the reasons for this discrepancy given the unknown composition of the candidature, it was widely reported that language teachers in Hong Kong adopted a product approach to the teaching of writing and tended to over-correct students'

written work and offer little in the way of explanations of errors (Pennington & Cheung, 1995; Tsui, 1996). As the Chief Examiner of the Writing Test in 2001 commented:

Candidates found this task hard, particularly 2B (explaining errors). Markers commented that the answers displayed a worrying lack of understanding of how English works. There was a tendency in 2B simply to describe the correction made in 2A without any attempt at generalisation or explanation. This suggests that many teachers have a restricted competence and lack awareness of the full range of English structures.

As this part of the Writing Test involved not just a certain level of language proficiency but also language and pedagogical content knowledge, candidates who had not received training in language found it difficult. However, the education authorities in Hong Kong believed that this innovative test was an assessment of a valid

Table 12.7 Design of LPATE speaking test

Test specification	Assessed skills/abilities/scales	Proficiency requirement determination method
Part 1: individual (10 min preparation + 5 min assessment) Task 1A reading aloud a prose passage Task 1B reading aloud a poem Task 1C telling a story/recounting an experience/presenting arguments	Part 1 <ul style="list-style-type: none"> • Pronunciation, stress and intonation • Reading aloud with meaning • Grammatical accuracy • Organisation and cohesion 	Candidates are assessed independently by two assessors who assign proficiency levels (1–5) according to descriptors on the six scales indicated. Candidates are assigned the means of these levels
Part 2: group interaction (10 min preparation + 10 min assessment) Candidates discuss language problems in a student's composition	Part 2 <ul style="list-style-type: none"> • Interacting with peers • Explaining language matters to peers 	

Table 12.8 Design of LPATE Classroom Language Assessment

Test Specification	Assessed skills/abilities/scales	Proficiency requirement determination method
Candidates were assessed on their language use during a live lesson in their school. Each candidate was visited twice by a different assessor each time. Each lesson consisted of one period (approx. 30–40 min) preceded by a briefing of 5–15 min, which was not assessed	<ul style="list-style-type: none"> • Grammatical accuracy • Pronunciation, stress and intonation • The language of interaction • The language of instruction 	Candidates were assessed as for the speaking test on the four scales indicated

requirement of English language teachers and helped to distinguish the LPATE from other, more generic English proficiency tests or assessments.

The Listening Test, in common with the Reading Test, proved to be reasonably straightforward to set, mark and standard set, though not necessarily to administer, given the complexities inherent in delivering all large-scale listening tests. The Test itself normally consisted of one long listening text, based on an authentic interview, which had been modified to fit the items constructed and re-recorded using voice artists. The difficulty came with delivery, though this was not specific to the LPATE. Since 1995, the HKEAA has utilised the services of Radio Television Hong Kong (RTHK) to broadcast the listening tests of the major public examinations, something that RTHK was rather reluctant to do by all accounts. This method relied on candidates bringing their own radio, and reception of the signal was often a problem, resulting in a large number of candidates requesting they be moved to the special room where the test recording was played through a standard CD player. For the LPATE (and the LPAT-Putonghua), given the smaller candidature, the recording was (and still is) played through a public address or loop system in examination centres.

During the original consultation process in which the different tests of the LPATE were developed and piloted, it was decided that there should be tests of Reading, Writing, Listening and Speaking as well as Classroom Language, to reflect the *target language use* (Bachman & Palmer, 1996) domains of English teachers in English teaching contexts. The decision to assess separately Speaking and Classroom Language was made for the following reasons:

Although Speaking is also assessed as one of the components of classroom language performance... it was decided to include an independent speaking test because teacher language performance has been observed to be different in the classroom compared to the language used among peers.... It is clearly not sufficient to judge the grammar, accuracy, pronunciation, stress and intonation of teachers merely on the language used in a lower secondary English language classroom. They must also be able to demonstrate a much higher level of language proficiency in other situations. It is important that the English language teacher is given an opportunity to demonstrate competence in a wide range of vocabulary and grammatical structures in professional settings. (Coniam & Falvey, 1998, p. 3)

In terms of test construct, the Speaking Test tasks were designed to replicate as far as possible the kinds of tasks that teachers needed to perform in their daily work. In Part 1 of the Test, candidates read aloud a prose passage and a poem and were assessed on their *Pronunciation, Stress and Intonation* and on their ability on the *Reading Aloud with Meaning* scale. A consistent issue throughout the development and administration of the Speaking Test was the legitimacy of including a poem as a text type, and indeed the need to include these tasks as candidates were also assessed on pronunciation, stress and intonation during another component of the LPATE, the Classroom Language Assessment. The decision to include these tasks in the Speaking Test was taken as it was considered that teachers of English must be able to serve as a model of English use for their students at the level of an educated Hong Kong user of English, and a legitimate means of achieving this is by reading aloud to the students. As candidates of the Assessment were free to choose the tasks that they performed during the Classroom Language Assessment and might well have avoided having

to read aloud to their class, it was felt necessary to include the reading aloud tasks in the Speaking Test. The choice of a poem as a text in addition to a prose passage was controversial but was found to be a good discriminator between candidates of differing abilities and provided assessors with the opportunity to measure each candidate's ability to deal with different types of text. Assessors were made aware of the fact that reading poetry aloud is a difficult skill, even for native speakers, and this was borne in mind during the assessor standardisation process and during the assessing itself. It was the hope of the test administrators that there would be some washback into English language classrooms with teachers taking the opportunity to read aloud more often to their classes. This issue is revisited in the next chapter.

The third task in Part 1 of the Speaking Test required candidates to speak on a given topic for a period of about two minutes, during which they were assessed on their *Grammatical Accuracy* and *Organisation and Cohesion*. It was the one task which required candidates to demonstrate the ability to construct discourse and present it accurately and cohesively to a potential audience, as they might need to do when presenting in English to colleagues. The main issue to emerge regarding this task was the choice of topic and whether sensitive topics should be avoided. The test developers felt that adult candidates and assessors should be able to deal with sensitive topics. However, feedback from assessors after the live administrations of the test indicated that they felt that it was unfair for some candidates to have to deal with such topics as death, illness, sex or other social issues while others might not have to do so. Given the high-stakes of the LPATE, when candidates were likely to be under great stress while taking the Test, it was considered that any extra source of stress should be avoided. To some extent, the test setters took the cultural characteristics of Hong Kong Chinese candidates into account when choosing topics and some topics that expatriate (i.e. Western) candidates may have had no problem discussing might have been considered taboo in Chinese culture. For this reason, the topics chosen were generally restricted to educational and language issues.

Part 2 of the Speaking Test brought candidates together in groups of three or four to discuss the errors in a student's composition and tested their ability to interact with colleagues in English in the kind of professional setting that might occur outside of the classroom. This part of the Speaking Test proved to be the least difficult for candidates in terms of the proportion of them achieving the required Level 3 on the five-point scale of proficiency on the two scales of *Interacting with Peers* and *Explaining Language Matters to Peers*. However, this did emerge as the part of the Test that concerned assessors the most as they had to assess each of the candidates in the group on the two scales as individuals, while at the same time considering the dynamics of the group. For example, the discussion may have been dominated by one candidate who was more confident than the others or was so eager to project themselves that they did not allow the others a chance to speak. In such cases, the assessor had to decide whether to intervene in the discussion to allow the other candidates a chance. While on the one hand it could be argued that part of the skill of interaction is to try to participate, on the other if a candidate is stopped from doing so, then it is not a fair assessment. While the task tried to simulate a real staff room discussion, it was an assessment and assessors had to assess a performance.

It was in Part 2 of the Speaking Test that candidates were tested on their ability to discuss the errors in a students' composition using appropriate metalanguage, and it was argued that this meant that within the Speaking Test candidates had to display not just speaking proficiency but also knowledge of the language of teaching. While competent English speakers could score highly on the first five scales on the test, without knowledge of the metalanguage and the ability to identify and discuss grammatical errors, they could not score well on *Explaining Language Matters to Peers*. It was this that perhaps made the LPATE Speaking Test unique as a test of language-specific oral proficiency. The Speaking Test was thorough in its demands on candidates and on assessors and normally took place over a five-day period. Each day consisted of two sessions, with two groups of four students being tested in each session, spread over three test centres. It required as great an effort of organisation to administer as it did to develop the tasks and train the assessors. Normally, between 40 and 50% of candidates attained Level 3 or above on the Speaking Test.

The Classroom Language Assessment was officially Paper 5 of the LPATE, though it differed from the other four papers in that it was conducted independently by the Education Department (later the Education Bureau) and could not be described as a 'paper' or a 'test' at all, rather it was (and still is) an assessment of teachers' use of English within their own classroom. For this reason, unlike the other components of the LPATE, the CLA could only be taken by serving teachers. The CLA attracted little attention maybe because the attainment rate for Level 3 or above was normally at around the 90% mark.

Sociopolitical Aspects of the LPAT

The LPATE proved to be a controversial initiative in its early days, evidenced by the extensive coverage given to it by the media. For an in-depth discussion of how the Hong Kong media have reported the LPATE, the reader is referred to Drave (Chap. 16, this volume). Opposition to the Assessment by some teachers, who saw it as an affront to their professionalism, emerged initially because of uncertainty over the ways in which it was going to be carried out, which teachers would need to sit for it and what the consequences would be if they did not (Coniam & Falvey, 1999). The Government's use of the threat of expulsion from the profession should teachers not be able to achieve the prerequisite standard angered teachers and led to calls for the LPATE to be boycotted, culminating in demonstrations against the Assessment, organised by the Hong Kong Professional Teachers' Union, in May and June 2000 (South China Morning Post, 28 May 2000; 11 March 2000). The current author, as Subject Officer, twice appeared on the Radio Television Hong Kong current affairs talk show *Back Chat* to respond to criticisms, mainly from the Professional Teachers' Union, that the Assessment was unfair and unreliable, and from members of the public that English teachers in Hong Kong lacked the necessary proficiency to teach. The related issue of public perceptions of the value of the LPATE as an

instrument of change in English language education in Hong Kong is explored by Drave in Chap. 17.

Yet despite reports in the press of unease among teachers, research showed that the majority of teachers supported it in principle. Of a total of 9179 teachers surveyed in 1996, 83% agreed or strongly agreed that there should be agreed on minimum standards of language ability for English language teaching purposes (Coniam & Falvey, 1996, 1999). In addition, support for the initiative was expressed by parents and members of the business community (Ho, 2000; South China Morning Post, 11 March 2000). Amid conflicting opinions, the Government tried to give the impression of being open to the needs of the public, but the public's needs and aspirations varied widely. Students, teachers, parents, principals, school governors, teacher trainers and others in the field all had different perspectives on the ways that education should be organised (Evans, Jones, Rusmin, & Cheung, 1998) and how and when new educational initiatives or reforms should be introduced. Governments often try to compromise and choose 'the path of least resistance' when deciding education policy, seeking, as Morris (1995) says, to preserve their status and, in some cases, social order. In a further relaxation of the Government's stance on the issue, the announcement was made in 2001 that proposals to eventually require all teachers who teach their subjects through the medium of English (and not only teachers of English or Putonghua) to reach a benchmark level of proficiency in the language would be abandoned (South China Morning Post, 4 June 2001).

Collective action by teachers played a large part in determining the development of the LPATE initiative. As Fullan (1991) argues, without the complete involvement of teachers, educational innovation is bound to fail. By showing teachers that it was prepared to listen to their concerns, the Government attempted to involve teachers more in the decision-making process. At the same time, the Government stood firm in its resolve to continue with the Assessment, with the aim of raising standards of language teaching, an aim which clearly had the support of most of the people of Hong Kong. Teachers did eventually become more accepting of the Assessment, as the increase in candidature and the reduction in press coverage of the results showed. As with any reform or innovation, a period of resistance is expected during the diffusion process before enough potential adopters of it, in this case the teachers, become persuaded to adopt it due to the factors within the innovation that facilitate change (Henrichsen, 1989; Rogers, 1995). Such factors relevant to the LPAT included

Relative advantage—improving standards.

Flexibility—Teachers could opt to do the Assessment or one of the training courses or a combination of the two, for which they received Government funding (see Mak & Xiao, this volume).

Trialability—Teachers were given a set time period to reach the required standard and so could take the Assessment as many times as they wished until they reached the required standard (although they only received funding for their first attempt).

Status—Teachers achieving the required proficiency level, who would previously have had no English teaching qualification, could gain *de facto* qualified English teacher status. In addition, teachers reaching overall Level 4 could be considered for promotion to head of department, and this feature (not officially sanctioned by

the Education Bureau) has resulted in the continued popularity of the LPATE and a perception that the teaching profession has been professionalised (see Coniam, Falvey & Xiao, Section IV, this volume).

In addition to the action of teachers leading to changes in the implementation details of the LPAT initiative, feedback from candidates led to modifications in the tests themselves. Examples of such modifications were in the Listening Test, in which more pauses were included and less writing in answers to allow candidates more time to process the input; and in the Speaking Test, in which the order of reading of poem and prose passage was reversed so that candidates read the passage first as it was considered less challenging as a first task in the test. However, these changes were little more than cosmetic and it became clear as the LPR deadline approached, that a more substantive revision would be needed.

Conclusions

The Language Proficiency Assessment for Teachers endured a number of teething problems during its development, from initial recommendation, through consultation, piloting and implementation to consolidation. It proved difficult at times to separate the technicalities of assessment from the sociopolitical baggage that the LPAT initiative carried. The Hong Kong Examinations and Assessment Authority and the Education Bureau were at the forefront of pushing through this innovative reform initiative. What has been shown is that given time and the cooperation and contribution of the major stakeholders, i.e. the teachers, the Assessment became a successful benchmark of language teachers' proficiency in the Hong Kong context.

As has been discussed in this chapter, there was initially strong opposition to the LPAT initiative, in particular to the LPATE. Experienced teachers of English felt threatened by having to prove themselves when in many cases they had been teaching English for many years. They felt that their professionalism and ability were being challenged and through representative organisations like the Professional Teachers' Union (PTU) demonstrated their objections to the initiative publicly, bringing the issue of English language standards to the forefront of public consciousness particularly during the period of 2000–2002. That this raised attention put pressure on the test developers is undoubted. It goes without saying that the HKEAA adopted (and still does) very careful, thorough and innovative test development procedures for every public examination, though it is also fair to say that the enhanced public attention paid to English language examinations in general and the LPATE, in particular, meant that the tests themselves were scrutinised for perceived faults in design or errors in production. In addition, the design of the LPATE itself caused some disconcertion as the 'minimum competence' model meant that candidates had to attain Level 3 on each component test, including the parts of the Writing and Speaking Tests that assessed subject knowledge.

The public attention paid to the LPATE faded as the Assessment and the LPR policy became more accepted and embedded within the local education system,

consistent with established conceptions of educational change. As the deadline for serving teachers to attain the LPR approached, questions began to be asked about the future of the LPATE, essentially, would it still be needed? The EDB felt that there was a need to retain the Assessment for new teachers of English as an option for schools who may need to reposition teachers who might have been trained to teach other subjects, and for those entering the profession directly from teacher education. The acceptance of the LPATE by schools meant that in many cases schools began to require an LPATE Level 3 (or even 4) even when novice teachers had already been subject trained. In addition, schools began the practice of requiring an LPATE Level 4 for promotion purposes, so that heads of department would need to show a higher proficiency level. While this ensured the continuation of the LPATE, it had become clear that EDB would need to revisit the design of the LPATE test components and carry out some research into it and enact necessary revisions so as to appease critics of the Assessment. This revision process will be described in the following chapter.

Notes

1. The term “benchmark” was used initially to refer to the language proficiency level or standard that the Government of Hong Kong wished language teachers to attain. This term became widely used to refer to the proficiency level(s) and the test or assessment itself and is still in use today, although the official name of the Assessment is the LPATE (English Language) or LPATE. (For a detailed discussion of benchmarks in language proficiency see Falvey and Coniam (1997) and Coniam and Falvey (1999). In this paper, “benchmark” will be used in the context of describing the development of the LPATE as it was the term in use at that time.
2. In terms of education in Hong Kong, the term *Chinese* refers to spoken Cantonese and Standard Written Chinese. Putonghua (or Mandarin) is taught as a separate subject in Hong Kong secondary schools.
3. Each public examination has a Subject Committee which oversees its operations, the membership of which is representative of the stakeholders involved. In the case of the LPATE, this meant school teachers, teacher educators, university professors, school principals and education officials.
4. It was acknowledged that this process had its limitations given that the tests equated with the anchor tests were the prototype versions rather than the final versions, but as the standard-setting took these results as advisory, it was considered acceptable. For more on the LPATE standard-setting procedures, Urmston (Chap. 13) and Drave (Chap. 14) of this volume.
5. Rasch measurement was used to equate the scores on the test with the scores on the anchor test of a sample group of test-takers. The Rasch analysis provided a common metric against which the performance of all test-takers could be mapped. From this, cut scores for the test were determined, as they aligned with those of the anchor test. For more on the LPATE standard-setting procedures, see Urmston (Chap. 14) and Drave (Chap. 15) of this volume.

6. A modified Angoff Method (Angoff 1971) was used whereby a panel of experts reviewed the test content to arrive at agreement on item difficulty and thereby determine the score that a minimally competent teacher should achieve on the test, i.e. the passing or benchmark score. This score was then assigned as the cut score for Level 3 on the 5-point scale of proficiency. The process was repeated to obtain cut scores for the other levels of the Assessment. This is explained more fully in Chaps. 14 and 15.

References

- Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement*. Washington D.C.: American Council on Education.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.
- Choi, C. C., & Lee, C. (2009). Developments of English language assessment in public examinations in Hong Kong. In L. Y. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner*. London: Routledge.
- Coniam, D., & Falvey, P. (1996). *Setting language benchmarks for English language teachers in Hong Kong secondary schools*. Hong Kong: Final report prepared for ACTEQ (The Advisory Committee on Teacher Education and Qualifications).
- Coniam, D., & Falvey, P. (1998). *Validating the classroom language assessment component: The Hong Kong English Language benchmarking initiative*. Hong Kong: Advisory Committee on Teacher Education and Qualifications.
- Coniam, D., & Falvey, P. (1999). Setting standards for teachers of English in Hong Kong—The teachers' perspective. *Curriculum Forum*, 8(2), 1–27.
- Coniam, D., & Falvey, P. (2013). Ten years on: The language proficiency assessment for teachers of English (LPATE). *Language Testing*, 30(1), 147–155.
- Evans, S., Jones, R., Rusmin, R. S., & Cheung, O. L. (1998). Three languages: One future. In M. C. Pennington (Ed.), *Language in Hong Kong at century's end* (pp. 391–418). Hong Kong: Hong Kong University Press.
- Falvey, P., & Coniam, D. (1997). Introducing English language benchmarks for Hong Kong teachers: A preliminary overview. *Curriculum Forum*, 6(2), 16–35.
- Fullan, M. G. (1991). *The new meaning of educational change*. New York: Teachers' College Press (With S. Stiegelbauer).
- Government of the Hong Kong Special Administrative Region. (2000). *Syllabus specifications for the language proficiency assessment for teachers (English Language)*. Hong Kong: Government Printer.
- Henrichsen, L. E. (1989). *Diffusion of innovations in English language teaching: The ELEC effort in Japan, 1956–1968*. New York: Greenwood Press.
- Ho, K. (2000, June 9). *Let's stand up for the benchmark tests*. Hong Kong iMail.
- Hong Kong Education Commission. (1995). *Education commission report no. 6*. Hong Kong: Government Printer.
- Morris, P. (1995). *The Hong Kong school curriculum: Development, issues and policies*. Hong Kong: Hong Kong University Press.
- Pennington, M. C., & Cheung, M. (1995). Factors shaping the introduction of process writing in Hong Kong secondary schools. *Language, Culture and Curriculum*, 8(1), 15–34.
- Rogers, E. M. (1995). *Diffusion of innovations* (4th ed.). New York: Free Press.
- South China Morning Post. (2000, May 28). *Teachers vow to boycott test*.
- South China Morning Post. (2000, March 11). 6000 march.

South China Morning Post. (2001, June 4). *Extension of language tests may be scrapped.*
Tsui, A. B. M. (1996). Learning how to teach ESL writing. In D. Freeman & J. C. Richards (Eds.),
Teacher learning in language teaching (pp. 97–120). Cambridge: Cambridge University Press.

Alan Urmston is an Assistant Professor in the English Language Centre at the Hong Kong Polytechnic University as well as teaching on undergraduate and postgraduate programmes; he coordinates assessments for the Centre. His main publication and research interests are in language assessment and sociolinguistics.

Chapter 13

The Revision of the LPATE



Alan Urmston

Abstract This chapter documents the revision of the Language Proficiency Assessment for Teachers (English Language) (LPATE). The chapter provides details of the first two stages of the revision project: the review of the LPATE papers and *Syllabus Specifications* (Government of the Hong Kong SAR, 2000); and the pilot tests and subsequent analysis of results carried out on the revisions proposed in the review. The body of the chapter consists of the details of the review and information concerning the procedures and methodologies used during the pilot tests as well as the results from the tests and the consequent analysis of the results required in order to both validate the new versions of the tests and to ensure consistency with the previous versions. The chapter closes with recommendations as to the implementation of the new versions of the tests based on the findings from the review and the pilot testing.

Introduction

In January 2005, the Education Bureau issued a Consultancy Brief entitled *Revision of the Language Proficiency Assessment for Teachers (LPAT) (English Language/Putonghua)*. This document stated that:

The LPAT, being a high-stakes assessment, has attracted concern from the public regarding candidates' performance in various LPAT papers since its first administration in 2001. In the course of post-mortem reviews of previous administrations of LPAT, the Subject Committees have made a number of recommendations for revising and improving the assessments for the consideration of the LPAT Main Committee. At its meeting on 9 December 2004, the Main Committee endorsed the proposal of revising the LPAT and agreed that:

A. Urmston (✉)
English Language Centre, Faculty of Humanities, The Hong Kong
Polytechnic University, Hung Hom, Kowloon, Hong Kong
e-mail: alan.urmston@polyu.edu.hk

- *Suitably qualified and experienced professional consultants should be engaged to undertake the revisions.*
- *As the LPAT has gained public recognition over time, it is important that any revisions made should preserve the form and face validity of the assessments.*
- *The revised LPAT is to be used after the 2005/2006 school year, by which time all serving teachers should have attained the LPR and the revised assessments would affect new teachers only.* Education Bureau (2005, p. 3)

The LPAT Main Committee [Note 1] therefore approved revisions to the LPAT (English Language) (LPATE) beyond the 2005–06 school year, and the HKSAR Government invited interested parties to submit proposals to select a suitable team to carry out the revisions. As a result of this process, a research team led by the current author was selected to carry out the revisions. In accordance with the agreement between the Government (represented by the LPAT Team) and the revision team, the first task of the revisions was to carry out a review of the existing LPATE and its Syllabus Specifications. Following the review, new test versions (for the Reading, Writing, Listening and Speaking Tests) as well as for the scales and descriptors (for CLA, and the Writing and Speaking Tests) were developed and piloted in early 2007.

This chapter first presents a description of the review of the LPATE and Syllabus Specifications leading to the recommendations of revisions to be carried out. The chapter then describes the development of pilot versions of the revised test papers, the processes involved in the pilot testing and the analysis of the results from the pilot testing. The chapter describes the recommendations made for revised testing and administrative procedures to enable the successful implementation of the revised test versions and outlines the work done in the final stage of the project, in which revised test specifications (for EDB/HKEAA use), together with *Candidate Guidelines*, would be produced.

Review of the LPATE and Syllabus Specifications

The two major parts of the review of the LPATE (the ‘Assessment’) and Syllabus Specifications were a survey of major stakeholders and an analysis of the tests/tasks administered during the period between 2001 and 2007, which consisted of a total of nine separate administrations [Note 2]. As a first stage in the review, the major stakeholder groups in the Assessment were identified as follows:

- Teacher candidates;
- School principals;
- Teacher educators;
- Test developers;
- Markers/assessors;
- Classroom Language Assessment (CLA) assessors
- LPATE Subject Officers (past and present).

Table 13.1 Focus group discussion participants

Participant type	Number
Teacher candidates	2
School principals	4
Teacher educators	5
Test developers	4
Test markers/Assessors 1	3
Test markers/Assessors 2	4
LPATE past and present Subject Officers	3
Total	25

To gather the views of these stakeholders on the LPATE and the potential ways in which it might be revised, it was decided to hold focus group discussions that would examine both the issues surrounding the LPATE that were specific to each group, and those that were of a more general nature. It was anticipated that the views of these stakeholders would serve as the basis for the consideration of revisions to the Assessment. A total of 25 participants attended seven different focus group discussions, as can be seen from Table 13.1.

All of the focus group discussions were audio recorded (with the participants' consent) so that a summary of the main points made in the discussion might be prepared.

Further, to gather the views of a larger sample of stakeholders, a questionnaire survey (see Appendix A “[LPATE Performance Descriptors: Writing \(Part 1: Composition\)](#)”), encompassing teachers and principals from approximately 100 primary and secondary schools, was carried out from December 2006 to February 2007. The results of this questionnaire study were used to further inform the review. The questionnaire was sent by mail to 100 schools in Hong Kong. The schools were selected to be a representative sample of all primary and secondary schools in terms of location, banding and medium of instruction (Chinese or English). A total of 77 completed questionnaires were returned—77% being a very high return rate for a postal survey without incentive (see, e.g. Blumberg, Fuller & Hare, 1974; Denscombe, 1998). In their returns, the majority of respondents (84%) were in agreement that the *Syllabus Specifications* should be revised.

In the second stage of the review, the team examined available data on the tests/tasks from the LPATE previously administered. These data included the previous tests/tasks themselves, including marking schemes and examiners' reports; candidate scores; item analysis of test/task results; and other information as provided by personnel from the Hong Kong Examinations and Assessment Authority (HKEAA) responsible for various aspects of the LPATE. One administration of the LPATE (March 2004) was selected to conduct a detailed analysis of candidate scores on the four different components of the Assessment. This administration was chosen for two reasons: there were more data available on this administration than on any of the others, and this administration had the highest candidature (3725). The data used and the analytical procedures carried out on them are described in Table 13.2.

The analysis of test/task data was carried out selectively with the purpose of clarifying/supporting the points raised during the focus group discussions. A sample

Table 13.2 Data analysis Procedures (from results of LPATE 2004 (March))

Data used	Procedures carried out	Purpose
<i>Reading</i>		
Raw scores of candidates on Parts 1 and 2	Correlation between Part 1 and Part 2	Determine relationship—Are they measuring similar traits?
	Correlations with other components	Determine relationships amongst test components
Question paper	Item content analysis	Describe construct being measured
<i>Listening</i>		
Raw scores of candidates	Correlations with other components	Determine relationships amongst test components
Question paper	Item content analysis	Describe construct being measured
<i>Writing</i>		
Raw scores of test takers on Parts 1 and 2	Correlations between the five scales of performance	Determine relationships amongst scales
	Correlations between Part 1 and Part 2	Determine relationships between writing parts
	Correlations between scales of performance on writing and other components	Determine relationships amongst scales and test components
	Many-Facet Rasch Analysis	Investigate how the tasks, scales of performance and markers vary from expected norms
Question paper	Item content analysis	Describe construct being measured
<i>Speaking</i>		
Raw scores of candidates	Correlations between scales of performance	Determine relationships amongst test components
	Many-Facet Rasch Analysis	Investigated how the tasks, scales of performance and markers vary from expected norms

of raw score data from approximately 400 candidates was used for this analysis. Where appropriate, findings from the analysis were incorporated into the discussion of the revisions made to each component of the Assessment to clarify or support points raised. When considering the points raised in the group discussions and the questionnaire study, the team looked at the evidence from the analysis of test/task data together with other sources of information, such as their own experience of the LPATE and their own professional judgements as experienced practitioners in English Language assessment. Resulting recommendations were put forward for consideration by the LPAT Team for the revision of the LPATE. The recommendations that were endorsed by the LPAT Team are labelled as **Revisions** and are presented as such, for example:

Revision: The Multiple-Choice Cloze component was to be removed from the Reading Test

Chapter Structure

The subsequent part of this chapter now examines and discusses each of the component papers of the LPATE in turn. In the analysis, the revisions endorsed by the LPAT Team are first presented together with supporting discussion. Details are then provided of the development of new versions of the component papers and the piloting of them.

The Pilot Tests

As a result of the review of the LPATE, recommendations were made as to how the tests and assessments should be revised and there ensued a period of test development and moderation in early 2007 in which new test specifications were drafted and the new test versions were piloted. In parallel with the development of the test versions was the revision of the scales and descriptors for the criterion referenced tests of Writing (Part 1), Speaking and CLA. During the development of the new versions, the recommendations from the review were supplemented by the extensive individual experience of the team members as Chief Examiners, Setters and Moderators of the LPATE, and in the case of the Project Manager, as Subject Officer at the HKEAA from 2001 to 2005.

Procedures

The Pilot Tests of the revised LPATE test components were conducted at the Hong Kong Polytechnic University (HK PolyU) and at the University of Hong Kong (HKU). In the Pilot Tests, test takers were required to take both the new versions and current or ‘old’ versions of the test components to establish comparability between the new and the existing versions [Note 2]. Details of the tests are shown in Table 13.3.

In the design of the Pilot Tests, the test takers for the tests of Reading, Writing and Listening conducted at HK PolyU were divided into three groups. Each group did two new components and two of the 2003 (March) components. It was not feasible for all test takers to do both versions of each of the three components in one day. In the tests conducted at HKU, the test takers took those components that it was felt needed extra numbers—the new version of the Reading, Writing and Listening Tests—plus the 2003 (March) version of the Listening Test. It was considered that the numbers were sufficient to carry out test data analysis with a reasonable degree of reliability. Details of the various analysis procedures carried out are given in the sections devoted to the individual components.

Table 13.3 Pilot tests

Test component	No. of test takers
Reading (Pilot)	72
Reading 2003 (live exam, March)	103
Writing (Part 1) (Pilot)	116
Writing (Part 1) 2003 (live exam, March)	52
Writing (Part 2) (Pilot)	116
Writing (Part 2) 2003 (live exam, March)	52
Listening (Pilot)	89
Listening 2003 (live exam, March)	66
Speaking (Pilot)	94
Reading (Pilot)	29
Writing (Part 1) (Pilot)	41
Writing (Part 2) (Pilot)	41
Listening (Pilot)	40
Listening 2003 (live exam, March)	26

Participants

Though the LPATE is an open assessment in the sense that the general public can enter provided that they meet the entry criteria, the target candidature is teachers (or prospective teachers) of English in primary and secondary schools in Hong Kong. With this in mind, the revision team attempted to recruit a combination of pre-service teachers from the Hong Kong Institute of Education (HKIEd—now the Education University of Hong Kong) as well as some serving teachers to take part in the pilot tests. It was felt that the HKIEd would provide both the most suitable test takers and test-taking venue for the tests. Given the likely change in the demography of the candidature of the LPATE after the deadline of the start the 2006–07 school year for all serving English teachers to have attained the Language Proficiency Requirement, to a majority of pre-service teachers, it was considered appropriate to use pre-service teachers as participants in the pilot testing.

The following sections discuss in detail the different components of the LPATE, outlining the development of the new versions, the piloting of them and the analysis of the results of the piloting leading to the validation of each component.

The Reading Test

In its original design, the LPATE Reading Test consisted of two sections: a multiple-choice cloze section and a reading comprehension section.

Part 1: Multiple-Choice Cloze

Though the focus group participants expressed little opinion on this section of the Reading Test, it was the view of the revision team that it should be removed. Multiple-choice (MC) cloze is not obviously a test of reading, and as such there is no rationale for its inclusion [Note 3]. It was not considered to be a good model of assessment for teachers, especially given the changes to the curriculum that are in progress. Amongst the shortcomings of this test method are that it requires the candidates to make choices amongst language presented to them, rather than to produce evidence of their ability to make sense of a text themselves, or to complete a text using their own active vocabulary or grammatical knowledge. Furthermore, it was felt that the standard-setting procedure would be more straightforward if this part of the paper were removed, leading to greater transparency and reliability [Note 4].

Revision: The Multiple-Choice Cloze component was to be removed from the Reading Test

The revision team conducted a content analysis of sample MC cloze components from previous administrations of the LPATE. The analysis suggested that the focus of items in this component was on:

- Lexical choice and knowledge, including collocation, colloquial and idiomatic expressions;
- Grammatical knowledge, including use of articles, prepositions, verb forms; and
- Cohesion, including conjunctions, lexical reference (pronouns, substitution).

Items covering these aspects of knowledge would be included in Part 2 (the reading comprehension component), as both MC and constructed response questions. Closed constructed response items can be used where there is only one possible answer, making marking simpler and more reliable, while still requiring candidates to supply the linguistic items themselves, rather than choosing from alternatives provided by the test developers. For example, candidates might need to complete a paraphrase of a sentence or idea in the text [Note 5].

Part 2: Reading Comprehension

In the early versions of the Reading Test, texts were chosen which adhered quite rigidly to the topic areas outlined in the Syllabus Specifications; i.e., they were teaching- or language-focused. Later, more varied, though related, topics were chosen. It was felt necessary to do this to retain a certain ‘freshness’. However, the moderation committee had been concerned about the relevance of certain texts to both primary and secondary focus candidates and had had to consider when choosing passages whether it would be reasonable for primary teachers to be reading such passages or articles. It was felt that while the skills required were applicable to both primary and secondary candidates, the platform, i.e. the texts, may not have been. This is related to what teachers actually do read rather than a perception of what they should read.

In general, it was felt that texts should be relevant to teachers and should not advantage or disadvantage any particular subgroup. However, at the same time, topics could be chosen that would be within the sphere of interest of language teachers, even if they were not specifically about language or teaching. The view of the revision team was that fundamentally texts should be relevant to teachers or relevant to the work teachers do within and outside the classroom. It was important to include a range of different topics in the Reading Test, rather than a single topic because of the cumulative effect on performance of using a single topic, which tends to undermine both the reliability and the validity of the test results, since those candidates to whom a topic is more familiar, interesting or accessible will gain increased advantage as the test progresses, while a corresponding disadvantage is suffered by those who find a topic less familiar, interesting or accessible.

Revision: A variety of unlinked topics were to be included in the Reading Test

If the MC cloze were to be removed, it made sense to include three texts in the Reading Comprehension section: one extended text on an education-related topic, and two shorter texts, on more general topics, and representing different genres. This would address many of the concerns raised in the focus group discussions (as described above), as well as those raised over the years by the LPATE Subject Committee, related to the need for the Test to include a range of topics as well as a range of genres.

Revision: The Reading Test was to include three texts with a total word length similar to the existing total of around 1500 words

As with the existing Reading Comprehension component, education-related texts were considered suitable for this part of the Test. One extended text would be related to education as this kind of reading is relevant to teachers' ongoing professional development. Topics and texts would also relate to those teachers are likely to encounter when working with students in the classroom. This would broaden significantly the range of topics to be included in the Reading Test, as many of them would be of a very general nature.

Revision: A variety of text types were to be included in the Reading Test

The revision team was of the opinion that it would be useful to retain one extended prose passage, as mentioned above, but other texts should be short, to ensure variety. In their daily work with students, teachers are required to use a variety of text types, and it is therefore important to represent variety, as far as possible in the Reading Test. Possible text types identified were:

- Narratives (relevant to literature, students' personal and creative writing, intensive reading schemes);
- Arguments (related to persuasive texts of all kinds, both written and read by teachers and students);
- Descriptions (relevant to many education-related materials);
- Dialogues (relevant to literature, written interviews);
- Explanations (represents the language of textbooks).

Some additional text types, such as procedures or reports, are relevant to teachers, but it was thought that these may be hard to incorporate, as it is often difficult to write good items based on such texts, given the limitations in reading skills or operations that readers need to engage in when reading such texts (or *item intents*—see below). For example, texts which are mainly factual in nature include little in the way of interpretation or inference. [For an in-depth treatment of the development of Reading Tests, the reader is directed to Alderson (2000) and Hughes (2003)].

Development and Piloting of the Reading Test

Based on the recommendations above, a new Reading Test was developed consisting of three reading texts with a word length of approximately 1600 words. The texts covered topics as varied as the energy use of large computer servers, the effect of the Chinese calendar on the birth rate and the problems faced by teachers in regard to bullying in schools. The revision team considered these topics to be (a) of interest to English teachers and/or (b) the kinds of texts that they would be likely to use in their teaching.

A list was produced of the skills/subskills being tested by each item on the paper so as to form a checklist to ensure that a range of skills was being tested. The term *subskills* may be a kind of shorthand for what the tests or tasks try to test. This term needs elaboration, as it is not just subskills that we need to describe. **Item intent** might be a more appropriate term and has been used previously in the context of labelling the what it is that items are designed to test (e.g. Filipi, 2012). This term has the advantage of expressing what the test or task is designed to test, rather than making a guess about how the candidate responds to tasks. The use of *item intent* allows for a description of what candidates at different levels are typically able to do. The item intent should try to include a description of features which contribute to an item's ease or difficulty.

Revision: More explicit guidance was to be provided for test developers in writing items, for example, by providing a description of 'item intents'—for test developers to be more confident that a sufficiently wide range of features of the construct of reading is tested in each paper produced.

For the pilot version of the Reading Test, a variety of test items was produced designed to attain an accurate and reliable measure of the test takers' reading comprehension. a list of item intents for this test is shown in Appendix B “[LPATE Performance Descriptors: Speaking](#)”. Most the items were dichotomously scored, while a small number were partial credit. A total of 101 test takers took the test, with the test papers marked by an experienced LPATE Reading Test marker who was also a member of the Reading Comprehension moderation committee. The results of the test were analysed using the one-parameter Rasch model, on the basis of which decisions were made as to which test items should be omitted from the test prior to standard setting. The results of the analysis of the test (after omission of unacceptable items) are shown in Table 13.4.

Table 13.4 Test results for the reading test

Number of items in initial version of test	62
Number of items (after item deletion)	52
Number of test takers	101
Maximum raw score	52
Mean	28.2 (54%)
Standard deviation	6.13 (12.3%)
Internal consistency (Cronbach's alpha)	0.72
Standard error of measurement	3.26 (6.5%)

The test results' analysis showed that with a mean percentage score of 54%, the test was of a suitable level of difficulty for the intended test takers (see Gronlund, 1985, p. 103). Test reliability—Cronbach's alpha—was 0.72, which was comparable to figures typically achieved for previous LPATE Reading Tests. The figure observed in the trial may have been affected by the relative homogeneity of the test taker sample (as evidenced by the relatively low standard deviation) as well as by the lack of stakes attached to the trial, for trial participants (they had engaged in no practice for the test, and their studies and careers would be unaffected by the results). The standard error of measurement was below 10%. See also Drave, Chap. 15, this volume.

Standard Setting

Maintaining a consistent standard across different versions of the Test is extremely important. This had been achieved previously by equating a new Test with an 'anchor' version of the Test and by Expert Judgement (modified Angoff procedure). The revision team considered this to be an appropriate method but flawed due to significant differences between the anchor test and the later versions of the paper. A more accurate process, which is analogous to the use of an anchor test, but with some important differences, was suggested. (See Eckes, 2009, for a practical guide to the use of anchor tests in standard setting.)

Revision: A different approach to standard setting was to be adopted for the Reading (and Listening) Tests using a single scale of items rather than a fixed anchor test.

A trial group (pre-test or post hoc) would complete two versions of the test. These would be analysed using Rasch measurement. The analysis would place all items from both versions onto the same (logit) scale. Previous Reading Tests could be used to establish comparability with past standards. A passing level on the (logit) scale would be set using present standard-setting procedures. Future versions of the paper would be equated to this scale by a similar process.

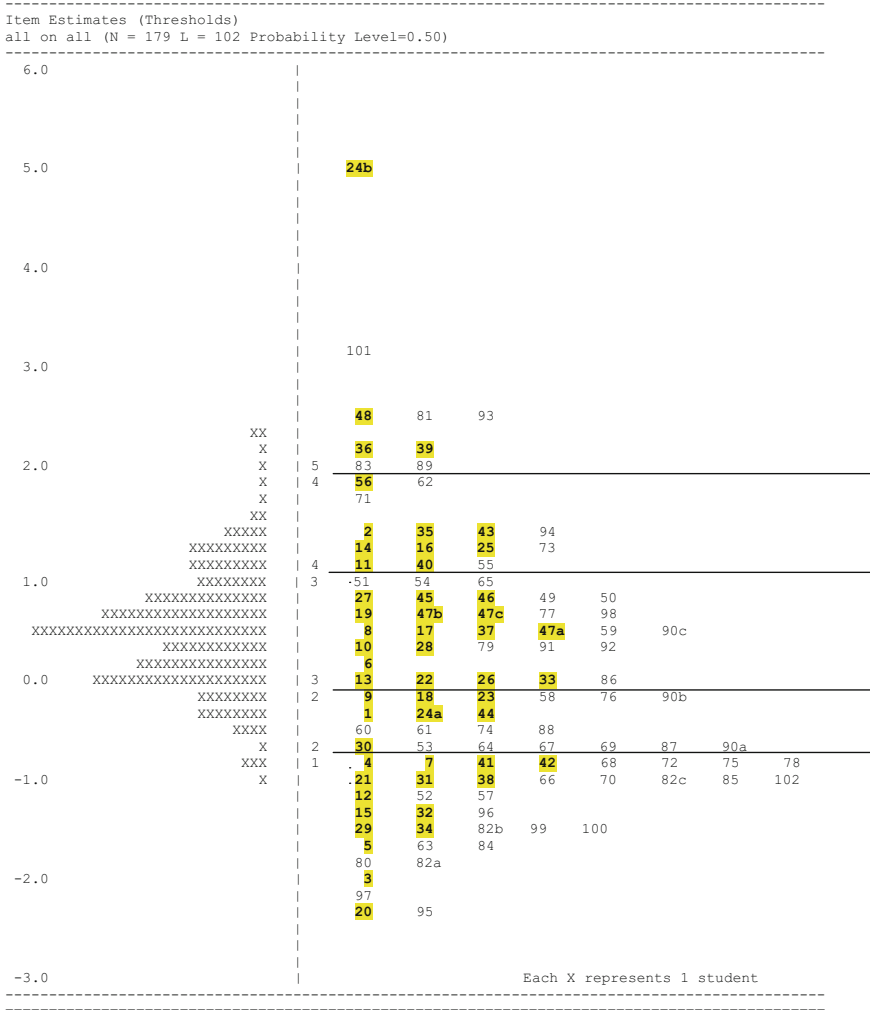


Fig. 13.1 Item map for the Reading Tests

To this end, the results from both the old version of the Reading Test [2003 (March)] and the new version were compared. As the Rasch model aligns items on a single scale of difficulty, it was possible to derive cut scores on the new version of the test by comparison with the cut scores on the old version. The results of the Rasch analysis of both versions of the Reading Test are shown in Fig. 13.1.

Figure 13.1 shows the item map for the Reading Tests. The items from both the test versions were plotted together. Items highlighted were from the new version of the Test, with partial credit items indicated (e.g. 24a, 24b). As the cut scores for the old version were already known, it was possible to draw lines across the map at the

Table 13.5 Cut scores for the reading test from Rasch analysis

Level	Score range suggested by Rasch analysis
5	48–52 (max.)
4	40–47
3	23–39
2	16–22
1	0–15

Table 13.6 Final cut scores for the reading test by Expert Judgement

Level	Score range provided by Expert Judgement
5	44–52 (max.)
4	38–43
3	24–37
2	17–23
1	0–16

points on the logit scale which corresponded to the boundaries between levels. The cut scores for the new version of the Test could then be extrapolated. For example, the cut between Level 2 and Level 3 for the old version was at 33/34 marks, so the number of items below the line dividing Level 2 and Level 3 should be at least 33 for the old version of the Test. In fact, it was 34, as there were three items at the same logit value. As error of measurement is taken into account during the Expert Judgement exercise, it is better to be conservative in the allocation of cut scores at this stage. The number of items on the new Test below this line added up to 22. Therefore, the cut score for Level 3 could be estimated to be at around 23 marks. The same procedure was carried out for the other proficiency levels, and the following cut scores for the new Reading Test were estimated (Table 13.5).

In accordance with the methods of standard setting employed by the HKEAA for the LPATE, these cut scores were taken into account by the members of a panel of Expert Judges when setting the standards for the new version of the Test. After completion of the Expert Judgement exercise, the final cut scores for the Reading Test were agreed upon and are shown in Table 13.6.

On the basis of the cut scores recommended by the Expert Judgement panel, the following is a breakdown of the results of the pilot test in terms of attainment of proficiency levels.

Table 13.7 shows that around 79% of the Pilot Test test takers attained Level 3 or above on the new version of the Reading Test.

Table 13.7 Proficiency levels in reading attained by pilot test test takers

Levels	5	4	3	2	1	3 or above
No. of test takers	2	6	72	18	3	80
Per cent (%)	2	6	71	18	3	79

Test Procedures

Pre-testing is essential for the Reading Test as item analysis allows objective decisions to be made on revision of the items before the paper is administered to its target population. In addition, when developing the Test, the marking process always has to be considered. Principles of marking need to be published so that candidates are aware of them.

The revision team agreed with the test developers that the existing procedures for test development adopted by the HKEAA worked well and should be retained.

Revision: The current procedures of test development including pre-testing were to be retained.

Summary of Revisions to the Reading Test

Based on the recommendations to emerge from the review of the LPATE, a new Reading Test was developed, piloted and standardised. The new Test consisted of 52 items. The Test had three sections of Reading Comprehension involving three separate reading passages and would serve as a sample for future administrations of the LPATE and as such would be included in revised *Guidelines for Candidates*.

The Listening Test

The consensus of opinion from stakeholders who took part in the review was that the Listening Test was both difficult and involved a complex variety of skills, and that it was not always clear what was being tested, though it was acknowledged that over time the paper had been improving. The inclusion of more speakers, for example, made for more interaction and hence was a more realistic test than previous (when there had been just two speakers—an interviewer and interviewee) and ought to discriminate better. In addition, efforts had been made to cut down on the amount of writing a candidate needed to do, though a balance had to be made between different item types for reliability.

It was also felt that there should be some ways of making the test more natural, such as by using video input rather than just audio as the present format had received many complaints that it was an unrealistic task (as all listening tests are). In addition, the topics chosen for the Listening Test, in a similar vein to the Reading Test, could be varied, so long as they did not advantage some candidates over others. In fact, a variety of topics is essential to obtaining reliable measurement of the trait of interest. In addition, using linked topics compounds any effect of background knowledge, candidate interest/motivation, etc., as the test continues. De-linking each section of the listening allows independent measurement of each candidate's ability to respond

to the items specific to each piece of stimulus, untainted by performance on previous items (or lack of concentration, or any other confounding, irrelevant factors).

Revision: Different (more than one) and unrelated topics were to be included in the Listening Test.

It was considered that the test should include three or more separate sections covering different topics. It should be possible to draw on sections from RTHK transcripts or recordings, for example [Note 6], to provide a suitable source of stimulus material from which listening texts may be developed. The change in the target test population, from serving teachers to pre-service teachers, made it feasible and appropriate to cover a much-expanded range of topics. As for the Reading Test, it was felt that candidates should be encouraged to listen to a variety of sources of English that they may well use in their teaching.

Some suggested topics for Listening were: the environment; animals; sport; health; transport; shopping; parks; holidays; fashion; pop music; TV; movies; mobile phones; technological developments—although care would need to be taken with regard to specialist vocabulary.

The question was asked as to why the recording was only played once, one that has been discussed frequently in research into the assessment of listening proficiency (e.g. Buck, 2001; Geranpayeh & Taylor, 2008). It was felt that there existed a trade-off between authenticity in listening to the recording once only and lack of authenticity in the nature of the tasks that had to be carried out. However, it was considered that this was still a fair and valid way of testing listening. Comments were also made about the speed of the recording and that in recent years it had appeared to get faster. In fact, the rate of delivery had not changed significantly over the life of the LPATE, suggesting that this was not an explanatory factor in differential candidate performance.

Revision: Features of listening such as topic, rate of speech, characteristics of speaker and item type were to be specified more explicitly in order to define more accurately the constructs being measured.

The specifications should cover:

- Number of listening texts (3 or 4 distinct texts);
- Variety of topics;
- Variety of speakers;
- Lengths of texts (increasing in length and perceived difficulty);
- Rate of delivery in words/minute;
- Interaction/text types (talkback, monologue/explanation, interview, conversation, etc.).

Participants considered that there should be clearer guidelines for the test developers to work from rather than relying on the intuition of the moderation committee, who had long experience with the test. As for the Reading Test, there should be a description of the intent of each item on the paper that the test developers could work with.

Development and Piloting of the Listening Test

The revised Listening Test as used in the Pilot Tests made use of three different spoken texts: an English Language learning webcast; a radio chat show on spoiled children; and a radio talk on American English. The revision team considered these topics to be (a) of interest to English teachers and/or (b) the kinds of spoken texts that they would be likely to use in their teaching. A variety of test items was produced designed to attain an accurate and reliable measure of the test takers' listening comprehension. A list of item intents for this test is shown in Appendix C "[Item Intents for the Revised Listening Test](#)". Most of the items (29) were dichotomously scored, on which test takers could score '1' or '0', while four of the items were partial credit, on which test takers could score '2', '1' or '0' marks. The total number of marks or maximum raw score on the test was 69. A total of 129 test takers took the test, with the papers marked by an experienced LPATE Listening Test marker. The results of the test were analysed via Rasch, and decisions were made as to which test items should be omitted from the final version of the Test. Two dichotomously scored items were omitted based on the facts that their item statistics (discrimination and facility indices) were below acceptable levels. In addition, three partial credit items were reduced from 2-point scoring items to dichotomous, 1-point scoring items for the same reason. The results of the analysis of the Listening Test (after omission of unacceptable items) are shown in Table 13.8.

The test results' analysis shows that with a mean percentage score of 56%, the test was of a suitable level of difficulty for the intended test takers, the mean score usually being in the 55–60% range. The internal consistency, which is a reliability figure equivalent to Cronbach's alpha, was 0.84, comparable to the reliability figures obtained from previous administrations of the LPATE Listening Test.

Table 13.8 Test results for the listening test

Number of items in initial version of Test	69
Number of items (after item deletion)	64
Number of test takers	129
Maximum raw score	64
Mean	36.0 (56%)
Standard deviation	8.73 (13.6%)
Internal consistency (Cronbach's alpha)	0.84
Standard error of measurement	3.49 (5.45%)

Table 13.9 Final cut scores for the listening test by Expert Judgement

Level	Score range provided by Expert Judgement
5	54–64
4	45–53
3	31–44
2	17–30
1	0–16

Standard Setting

The revision team adopted the same standard setting procedures as for the Reading Test, and the results, after Expert Judgement, are shown in Table 13.9.

Summary of Revisions to Listening Test

Based on the recommendations to emerge from the review of the LPATE, a new Listening Test was developed, piloted and standardised. The new Test consisted of 64 items. The Test had three sections of Listening Comprehension involving three separate spoken texts. The Test would serve as a sample for future administrations of the LPATE and would be included in forthcoming *Guidelines for Candidates*.

The Writing Test (Part 1)

Previously, in Part 1 of the Writing Test, candidates had to write an expository text. The view of the revision team was that the expository text had become somewhat formulaic or predictable and that it was necessary to vary the text types required.

Revision: Candidates were to continue to write one text in Part 1 of the Writing Test, but the text should no longer be restricted to expository only.

As with the Reading Test, the texts that candidates were required to write should have reflected the kinds of texts that they might encounter as a teacher. This should not be limited to those that they may need to write themselves, but also to those that they may require their students to write. The ability to perform themselves the tasks that they set for their students should be a requirement for a teacher. Possible text types identified were:

- Narrative;
- Description;
- Explanation;
- Procedure;
- Report;

- Mixed genre (e.g. a letter, complaining about something, involving recount and argument, resolution of a problem).

The revision team also believed that varying the text types required would have a beneficial washback effect on teachers as they would no longer simply practise one type of writing to prepare for the LPATE.

Development and Piloting of the Writing (Part 1) Test and Scales and Descriptors

In response to the above recommendations, a new writing task was developed, requiring candidates to write a descriptive/discursive text on relationships. This task was piloted on 157 test takers, and their scripts were double marked by experienced LPATE markers.

The participants in the focus group discussions pointed out that there was a need to revise the *Scales and Descriptors for Writing* due to the overlap between Scale 1 (*Organisation and Coherence*) and Scale 3 (*Task Completion*). This overlap meant that for Scale 3, there had to be a set number of clearly defined tasks that the candidates were asked to address and how well they did this was then made the basis for the score given for this scale.

Writing Part 1 also did not discriminate particularly well. The Many-Facet Rasch Analysis of the LPATE 2004 (March) data showed there to be 2.98 separate levels of ability (reliability of this separation = 0.90). With five levels of performance, this was clearly of concern. What it indicated was that markers worked mostly within the Levels 2–4, seldom awarding Levels 1 or 5. This may have been a reflection of the candidature, or it may have indicated that attention needed to be paid to the descriptors with a view to working with markers to see how they might be modified to achieve greater separation and hence reliability of measurement.

Revision: The scales and descriptors for Part 1 of the Writing Test were to be revised.

In accordance with this recommendation, the *Scales and Descriptors for Writing* were revised to take into account the concerns expressed by stakeholders during the review. The main changes can be summarised as:

- The removal of the ‘overlap’ between the scales of *Organisation and Cohesion* and *Task Completion* with regard to sensitivity to the text and audience. The criterion *Tone and Style* was to be included in Task Completion.
- The addition of *Lexical Accuracy and Range* to the *Grammatical Accuracy* scale to account for use of vocabulary.
- The change of format from a summary description paragraph to bullet-point form for ease of use by markers.

During the marking process, the scales and descriptors were amended to take into account the opinions of the markers.

Table 13.10 Measure of difference in test versions of writing (Part 1) tests

	Fair Average	Measure (logits)	Model error	Infit mean square
New test	2.9	-0.04	0.08	1.0
Old test	2.9	+0.04	00.18	0.8
Mean	2.9	0.00	0.13	0.9
SD	0.0	0.04	0.05	0.1

RMSE 0.14 Adj S.D. 0.00 Separation 0.00 Reliability 0.00
 Fixed (all same) chi-square: 0.2 d.f.: 1 significance: 0.65

In order to answer the question of whether the new task and the revised scales and descriptors were comparable in level of difficulty as well as in the constructs of writing that they were describing, a Many-Facet Rasch Analysis of the Pilot Test results was conducted. In the analysis, the results of both versions of the Writing Test, the new version and the 2003 (March) version, were analysed together. The FACETS output provided information on the Rasch model of the data that had been input and allowed for a single dimension on which the facets of test taker ability, marker behaviour (relative harshness or leniency) and task and item difficulty could be calibrated. In the case of the Writing Test, 'item' refers to the scores on the criteria of *Organisation and Cohesion*, *Grammatical and Lexical Accuracy* and *Task Completion*.

FACETS allowed comparisons of the two versions of the Test to be made with a view to seeing to what extent the two tests were measuring the same abilities in the same way. Table 13.10 shows the FACETS output for the two test versions.

Table 13.10 shows that there was no significant difference in the two test versions, with the new version recording a logit measure of difficulty of -0.04 compared to that of the old version of 0.04. This means that the test takers on average found it as easy to score on the new test version as on the old. The analysis also showed the model data fit of each test to be close to 1.0 and that the tests were therefore performing well.

FACETS also allows for comparisons to be made of the differences in the items being measured, in this case, the criteria of *Organisation and Coherence*, *Grammatical and Lexical Accuracy and Range* and *Task Completion*. Table 13.11 shows the analysis.

Table 13.11 shows that the criterion of *Grammatical and Lexical Accuracy and Range* proved to be the most difficult for test takers to score well on with logit measures of 1.11 for the new test and 0.75 for the old test. This is consistent with observations of the ways that candidates typically perform on the LPATE Writing Test as well as on most second language writing tests. To put it another way, raters, especially experienced ones (as these were), tend to rate language accuracy more harshly than other criteria (Barkouai, 2010). The other criteria were close in terms of measure across the two test versions, but as the model data fit for each criterion was within the range of 0.5–1.5 logits, it was considered that these criteria were performing as they would be expected to. Hence, the conclusion to be drawn was

Table 13.11 Measure of differences in testing criteria

	Fair Average	Measure (logits)	Model error	Infit mean square
OC (New test)	3.0	-0.39	0.13	0.8
GA (New test)	2.7	+1.11	0.14	1.1
TC (New test)	3.0	-0.86	0.13	1.2
OC (Old test)	2.9	-0.16	0.31	0.6
GA (Old test)	2.8	+0.75	0.30	0.9
TC (Old test)	3.0	-0.45	0.31	0.7
Mean	2.9	0.00	0.22	0.9
S.D.	0.1	0.70	0.09	0.2

RMSE 0.23 Adj S.D. 0.66 Separation 2.80 Reliability 0.89

Fixed (all same) chi-square:127.8 d.f.:5 significance: 0.00

Random (normal) chi-square: 5.2 d.f.:4 significance: 0.27

Key: OC = Organisation and Coherence; GA = Grammatical and Lexical Accuracy and Range; TC = Task Completion

that the revised Writing (Part 1) Test was equivalent in terms of constructs being measured to the previous versions.

It was decided to use the revised *Scales and Descriptors for Writing* for marking of both the old and new test versions as fundamentally the standard is set by the test setters and by the markers. If the markers found the scales and descriptors inclusive of the constructs that they were familiar with as LPATE markers and which they saw evidence of in the test taker's writing, and they found the scales and descriptors easy to use, then that would make for reliable marking. The revision team worked closely with the team of LPATE markers on the development of the scales and descriptors, and these were modified throughout the process of marking until the final version was agreed upon. In terms of equivalence with the existing scales and descriptors, for those test takers who took both the old and new tests, the following score correlations were found:

Table 13.12 shows that the scores across the two test versions correlated reasonably well (and were statistically significant) considering the small sample size and taking into consideration the improved scale content of the revised scales and descriptors and the different ways that the constructs had been defined. To conclude, assuming the markers of future administrations of the LPATE remained the same and went through

Table 13.12 Correlations between scores for test takers who took both writing tests ($n = 29$)

	Organisation and Coherence	Grammatical and Lexical Accuracy and Range	Task Completion	Test Mean
Correlations	0.49* ($p = 0.000$)	0.56* ($p = 0.000$)	0.36* ($p = 0.000$)	0.57* ($p = 0.001$)

*Correlation significant at 0.01 level (two-tailed)

similar training and standardisation procedures, the revised scales and descriptors should provide for more accurate and reliable measurement of writing ability.

The revision team recommended that the scales and descriptors in their revised form should not be provided for candidates as they were designed to be used by testers and may have been open to misinterpretation. Instead, a modified and simplified version was provided for candidates.

Test Result Reporting

To gain a better picture of the performance of a particular test version, the FACETS output is most easily interpreted by viewing the ‘All-Facet Vertical Rulers’ on which all the elements involved in the test—test takers, markers, criteria and scores—are aligned along the same logit scale. To illustrate this, the All-Facet Vertical Ruler for the revised Writing (Part 1) Test is shown in Fig. 13.2.

Figure 13.2 shows the distribution of test takers along the scale of ability in the second column, with each * representing 2 test takers. The markers are shown in the third column and it can be seen that they varied in degrees of leniency, with Marker E the most harsh (1.75 logits) and Marker F the most lenient (−2.38 logits). Under normal operational circumstances, more training and standardisation would be required, perhaps even removing the outliers (Raters E and F), though differences can be accounted for when using MFRA as the test taker receives a ‘Fair Average’ score rather than an ‘Observed Average’, the latter being the raw score. Any variation in the facets of marker, task or criteria are modelled by the programme and compensated for. (For a full discussion of the use of raw scores and Rasch-generated Fair Average scores in test result reporting, see Coniam (2008a, b).) With the current practice of reporting the mean raw score given by two markers, there is no way to account for differences in marking harshness in particular, other than the processes of marker training and check marking that were carried out at the time. It was the opinion of the revision team that the LPATE should adopt the use of Rasch modelling and Fair Average scores to both aid the process of test analysis and in the reporting of test scores.

Recommendation: Test results for the LPATE Writing (Part 1) Test were to be reported as a Fair Average score using MFRA rather than the mean score given by two markers.

It was possible to report a Fair Average score for each criterion of writing and an overall Fair Average score for the Test. This would contrast with the then existing process of reporting the mean scores given by each marker on the separate scales of performance and requiring the candidate to attain a Level 3 or above on each scale in order to reach the Language Proficiency Requirement. The effect of using Fair Average rather than observed or raw scores would be to reduce the number of false negative and false positive results for individual candidates, rather than to have a dramatic effect on the overall proportions of those passing or failing the Test. To investigate how the reporting of scores using the two methods might affect the results

Measure	Test taker	Test	Marker	Criteria	Scale
+ 10	*.	+	+	+	+(5)
+ 9	+	+	+	+	+
+ 8	*.	+	+	+	-----
+ 7	.	+	+	+	+
+ 6	*. **	+	+	+	4
+ 5	*. **.	+	+	+	+
+ 4	*. ***	+	+	+	+
+ 3	****. .	+	+	+	-----
+ 2	**. .	+	+	+	+
+ 1	**. *. *****.	+	+	GC	3
* 0	*. **. *****.	* T1	* D	* OC TC	* *
+ -1	.* *****	+	+	+	+
+ -2	***. **.	+	+	+	-----
+ -3	*. *.	+	+	+	+
+ -4	*. .	+	+	+	+
+ -5	+	+	+	+	+
+ -6	.*	+	+	+	2 +(1)
Measure	* = 2	Task	Marker	Criteria	Scale

Key: OC = Organisation and Coherence
 GA = Grammatical and Lexical Accuracy and Range
 TC = Task Completion

Fig. 13.2 Summary measure of all facets on Writing (Part 1) Test

that the candidates received, the results from the Writing (Part 1) Test were processed using FACETS to obtain Fair Average scores. The results are shown in Table 13.13.

Table 13.13 shows the numbers of test takers attaining the various levels for the Writing (Part 1) Test when calculated by the different methods [Note 7]. On a purely observed average of two markers' scores and requiring test takers to attain Level 3 or above on each scale, 43% would have attained the required proficiency level. If a Fair Average on each scale were used, then this figure would have increased slightly to 46%. If, however, the Fair Average scores on all three criteria had been used, approximately 73% of the test takers would have obtained Level 3 (on the basis of 2.8 or above) or 52% (on the basis of 3.0 or above). This shows that it was the policy of minimum competence on each criterion that had penalised candidates and kept the attainment rate down, rather than the actual ability of the candidates, an issue discussed by Coniam and Falvey (2001).

In practice, with larger test populations, greater numbers of markers and possibly different tasks, the use of Fair Average scoring would allow a more balanced and fairer assessment of candidates' proficiency. Further, the existing practice of allowing candidates one Level 2.5 providing all other scores were at least at Level 3 to attain the LPR would no longer be necessary as allowance was made for error within the Fair Average reporting.

Table 13.13 Comparison of numbers of test takers attaining different levels on writing (Part 1) test when calculated by Observed or Fair Average (*n* = 157)

Levels	5	4.5	4	3.5	3	2.5	2	1.5	1	3 or above
OC										
Observed Average	0	7	12	29	67	29	12	0	1	115 (73%)
Fair Average	0	6	11	26	73	25	14	0	1	116 (74%)
GA										
Observed Average	0	5	9	15	53	35	38	1	1	82 (52%)
Fair Average	0	5	12	10	72	17	40	0	1	99 (63%)
TC										
Observed Average	3	7	24	31	47	22	21	0	1	112 (71%)
Fair Average	3	6	26	27	50	23	21	0	1	122 (77%)
Overall										
Observed Average										68 (43%)
Fair Average										73 (46%)

Key: OC = Organisation and Coherence; GA = Grammatical and Lexical Accuracy and Range; TC = Task Completion

Summary of Revisions to Writing (Part 1) Test

Based on the recommendations to emerge from the review of the LPATE, a new Writing (Part 1) Test was developed and piloted. The results of the test were analysed and measures to ensure the fair reporting of candidates' performances recommended. The new test would serve as a sample for future administrations of the LPATE and would be included in forthcoming *Guidelines for Candidates*. The revised *Scales and Descriptors for Writing* would be included in the LPATE Specifications for use by test developers and markers. A simplified version would be provided for candidates.

The Writing Test (Part 2)

Perhaps unsurprisingly, it was Part 2 of the Writing Test that generated the most discussion during the review process. The focus group participants were unclear as to whether teachers' professional knowledge should be tested. On the one hand, this part of the LPATE tested the application of language to teaching-specific genres, but on the other it was felt to be too far removed from language proficiency to be included. Some of the typical comments on this part of the test were as follows:

- *Though explaining errors is part of the skills required of a teacher, should they be tested here?*
- *What is being tested, is it the knowledge or the ability to express that knowledge?*
- *The fact is that candidates have now had many years to learn how to do the tasks but the attainment rates have not improved indicates some problems with the validity of the tasks.*
- *How often do primary or secondary teachers need to write explanations of errors?*
- *You can't have a public test that has a pass rate of 30–40 per cent.*
- *The question is of whether Task 2A is a reading rather than a writing skill. Correction is usually not difficult for candidates.*
- *The scales and descriptors for Task 2 are not used as the nature of the tasks requires the use of a marking scheme and set cut scores.*

Participants agreed that Task 2A (Error Correction) contained a reasonable choice of item types and should be retained since the Correction of Errors is a necessary skill for language teachers.

Revision: Task 2A of the Writing Test was to be retained.

Error correction was considered a necessary part of English teachers' daily work, and the penalisation of incorrect changes may have had a beneficial washback effect, in signalling to teachers the need to exercise discretion in their use of the red pen.

The stakeholders felt unsure, however, as to whether it was proficiency that was being tested in Task 2B (Error Explanation) or capacity or content knowledge. There was some doubt as to whether teachers should have such knowledge and whether the LPATE should test it. As provision for this was now available elsewhere in terms of teacher education and subject knowledge courses, it may not have been necessary to include it any more. Some of the stakeholders even felt that the metalanguage involved in Task 2B of the Test was to a certain degree less advanced than that taught in teacher education programmes, so questioned the need to test it. No evidence was produced for this assertion.

For Part 2 of the Writing Test as a whole, MFRA of performance levels attained on a previous version [2003 (March)] showed there to be only 2.53 separate levels of ability (reliability of separation = 0.87). The error terms were large, mostly over one logit. The conclusion was that this part of the Test was poor at discriminating amongst candidates, partly because each candidate was scored on only two features of language (scoring categories) and, more importantly, a single task that could be analysed. A clearer picture of the reliability of this task would be obtained if individual items on this component were analysed, in the same way as was done in the Reading and Listening Tests.

The analysis of previous test/task data showed a correlation of 0.32 between Writing Part 1 mean score (aggregate across the three scoring criteria) and Task 2B. This is very low for two components of a language proficiency test, especially two components which both aim to address the same part of the construct, in this case, writing. Correspondingly, the correlation between Part 1 mean score and Task 2A was 0.49, a much higher figure. Taken together, these figures suggest that error correction represents a rather different aspect of writing.

It was therefore felt that a different approach was needed if candidates' knowledge of language were to be tested. The more objective Task 2A had proven to be a satisfactory task for measuring candidates' ability to recognise and correct errors. However, Task 2B, where markers needed to interpret the explanations of candidates, had proven to be less reliable, as markers frequently had difficulty understanding the candidates' explanations due to other factors (e.g. lack of coherence). This meant that it was unclear as to whether candidates' ability to explain errors was being judged. It was therefore suggested that Task 2B become a more objective type of task, requiring less interpretation from markers, with short answer items designed to measure candidates' knowledge of English.

Revision: Task 2B of the Writing Test was to be redesigned incorporating more objective items such as short answer or gap-fill.

The revision team did not entirely agree with the participants in the focus group discussions that the future candidature of the LPATE would already have been tested in areas of metalinguistic knowledge and the ability to explain errors to students and felt that it was necessary to retain a task that tested these areas.

An additional advantage of this type of task would be that the two tasks in Part 2 would become de-linked, allowing for more reliable measurement. Also, it would now be possible to use the item-based nature of Part 2 to carry out Rasch-based standard setting as outlined for the Reading and Listening Tests. This would help maintain the same standard across different versions of the Test.

Revision: The link between Task 2A and 2B of the Writing Test was to be removed.

This section of the Test was item-based, and as such a marking scheme was used. The design of the marking scheme had gone through a series of refinements to enhance the reliability of the marking such that the old LPATE Writing Scales 4 and 5 and the descriptors that went with them were no longer used.

Revision: The scales and descriptors for Part 2 of the Writing Test were to be removed.

A description of the required standard could be produced by describing the item types that candidates at that level are typically able to answer correctly. Informing candidates that a revised, more objective procedure would be used to set cut scores, considering the relative difficulty of the particular set of items included on the version taken by each candidate would enhance the face validity of this part of the LPATE.

Development and Piloting of the Writing (Part 2) Test

When designing the new version of the Writing (Part 2) Test, the revision team followed the recommendations listed above. The test consisted of a student's composition with numbered items as before, however, instead of the test takers being required to correct all items and then explain some of the same items, in the new version they had to correct the first nine items and then explain the remaining nine. (In each task, answers to the first of the ten items were given as examples.) This removed the link between the two tasks which had been a major criticism of Part 2 of the Test. In addition, in the new Task 2B, the candidates would need to complete the explanations of the errors/problems by filling in blanks in the explanations. A major criticism to emerge from the review was the difficulty that markers had in interpreting the explanations given by candidates. In addition, it had been a common complaint from candidates that the requirements for the explanation of errors/problems were unclear. The revised format made the requirements clearer and allowed for easier and therefore more reliable marking.

The 157 completed papers from the Pilot Tests were marked by experienced LPATE markers, and during the marking process, possible answers to every item were considered and the marking scheme expanded and refined. A variety of test items was produced designed to attain an accurate and reliable measure of the test takers' ability to recognise, correct and explain errors/problems. A list of the item intents for this task is shown in Appendix D "[Item Intents for the Revised Writing \(Part 2\) Test](#)". All items were dichotomous 'right' or 'wrong' items on which test takers could score '1' or '0' marks. The total number of marks or maximum raw

Table 13.14 Test results for the writing (Part 2) test

	Task A	Task B	Overall
Number of items	9	9	18
Maximum raw score	20	20	40
Mean	11.7 (58%)	10.8 (54%)	22.5 (56%)
Standard deviation	2.6 (13%)	3.6 (18%)	5.3 (13%)
Internal consistency	0.60	0.70	0.75
Standard error of measurement	1.60 (8%)	1.98 (10%)	3.58 (9%)

score on each task was 20. The results of the test were analysed using the Rasch software, enabling decisions to be made as to which test items should be omitted from the test prior to standard setting. In fact, it was decided that no items should be omitted, though a small number of items were modified for clarity to enhance their reliability prior to the final version of the Test being published. The results of the analysis of the Writing (Part 2) Test are shown in Table 13.14.

The test result analysis showed that with mean scores of 58% on Task A and 54% on Task B, the test was of a suitable level of difficulty for the intended test takers. (The mean score on an LPATE Writing (Part 2) Test was typically in the 60–70% range for Task A and 55–60% for Task B.) The internal consistency (reliability) figures of 0.60 for Task A and 0.70 for Task B were relatively low and possibly affected by the low number of items on each task and the relative homogeneity of the test taker sample [Note 8]. The reliability of these tasks was improved through modification of some of the items before the test was published. In addition, it shows that pre-testing of this part of the Writing Test was essential. The standard error of measurement was again acceptable, being below 10%.

Standard Setting

It was proposed that in future administrations of the LPATE, the standard for each new version of the Writing (Part 2) Test be set through a process of Expert Judgement, using procedures similar to those used for the Reading and Listening Tests. This was the only way in which the standard could be maintained across different versions and different administrations of the Test. To do this, the results from both the old version of the Writing (Part 2) Test [2003 (March)] and the new version were compared. Using the Rasch model to align items on a single scale of difficulty, cut scores on the new version of the Test were derived by comparison with the existing, fixed cut scores on the old version [Note 9].

After completion of the Expert Judgement exercise, the final cut scores for the Writing (Part 2) Test were derived and are shown in Tables 13.15 and 13.16.

Table 13.15 Final cut scores for writing (Part 2) Task A by Expert Judgement

Level	Score range provided by Expert Judgement
5	19–20
4	16–18
3	13–15
2	8–12
1	0–7

Table 13.16 Final cut scores for writing (Part 2) Task B by Expert Judgement

Level	Score range provided by Expert Judgement
5	18–20
4	15–17
3	12–14
2	8–11
1	0–7

Summary of Revisions to Writing (Part 2) Test

Based on the recommendations from the review, a new Writing (Part 2) Test was developed, piloted and standardised. The new Test consisted of 18 items with a total raw mark score of 40. The Test had two sections: Correction of Errors/Problems; and Completion of Explanation of Errors/Problems in a Student's Composition. The Test would serve as a sample for future administrations of the LPATE and would be included in forthcoming *Guidelines for Candidates*.

The Speaking Test

Tasks 1A and 1B—Reading Aloud (Prose and Poetry)

The majority view of the stakeholders who took part in the review was that the inclusion of the poem in Part 1 of the Speaking Test was frequently discouraging to candidates and should be removed. Sometimes, the poems were difficult to interpret and to read, and sometimes not representative of the kinds of poems that are used in classrooms. Despite this potential drawback, it was felt that reading aloud is a relevant skill for teachers (the *big book approach* requires it for primary teachers as does training students for the Hong Kong Speech Festival for secondary), and they may not manage it very competently. It therefore had relevance in this Test, by sending a message to teachers that this kind of language skill is valued. While reading a prose passage is relevant, response to poetry is very personal and its interpretation requires higher-order skills. Poems have multiple meanings, and assessment of the understanding of a poem would be necessary before deciding how to read it. Another

text type could be chosen, or it might be possible to have one text instead of two. This would impact on the third task in Part 1 (Task 1C), making it easier to control the timing.

The remark was made also that it was becoming more difficult to find suitable poems, especially as they could not be used more than once. As each administration of the Speaking Test required up to 12 different versions of the test and therefore 12 different poems, and this had to be done twice each year (between 2003 and 2005), it was proving to be a challenge for test setters.

Revision: The reading of a poem was to be removed from the Reading Aloud section of the Speaking Test.

Removing the reading of a poem would leave the reading of a prose passage only. It was considered that provided there was a sufficient sample for the assessors to make a judgement, then the task was satisfactory. It was recommended to reduce the length of the reading aloud task, so it could be completed within the 60s. It was also recommended that the content of the reading aloud passage be accessible in meaning to the candidates and reflect the level of language they could expect to actually be reading aloud in their classes. It was considered that texts of a literary genre, such as sections of narratives, descriptions, dialogues, etc., were most suitable as these were the kinds of texts that teachers read to classes and were consistent with the language arts approach promoted in the curriculum for schools.

A paragraph of 90–100 words would be expected to provide a sufficient sample for assessors to assess the features of interest in this task: pronunciation, intonation, stress patterns and pausing. Some of the tasks used in the past included only a limited range of sentence forms. It would be useful to include question forms (to allow assessment of rising intonation, for example); information in parentheses (to allow assessment of ability to indicate supplementary information, for example); and shorter sentences.

Task 1C—Telling a Story/Recounting an Experience/Presenting Arguments

Most stakeholders agreed that Task 1C was a valuable task and should be retained. The only issue raised regarded timing, and it was felt that it should be fixed, rather than being influenced by the amount of time the candidates took for Tasks 1A and 1B. In the past, candidates had sometimes had insufficient time to complete the task, making it difficult for assessors to assess features of the discourse such as coherence and organisation. It was considered essential that sufficient time was made available for all candidates to complete the task (taking into consideration the logistical considerations involved in test administration).

Revision: The allotted time for Task 1C in the Speaking Test was to be standardised.

It was felt that this task should be retained because it was necessary to include one task where candidates are required to produce an extended piece of discourse, in which their ability to organise content coherently, as well as associated language skills, could be assessed. The specifications should include a requirement for this task to draw on the ability to demonstrate higher-order language proficiency: providing relevant information, coherent structure and abstract and reflective language rather than features such as ‘safe’ concrete descriptions or recounts, or predictable topics.

A wide spread of assessor severity was observed in the Speaking Test data analysis. In order to improve measurement of candidates’ language proficiency, in addition to ensuring that rater training and standardisation was done properly, it was recommended that the revised Test makes the rating task easier for assessors to manage and the above measures were designed to achieve that. A further measure was to remove the thematic link between the tasks of Part 1. This would hugely increase the range of texts and tasks available, allowing far more flexibility to test developers.

Revision: The thematic link between tasks in Part 1 of the Speaking Test was to be removed.

As with all components of the LPATE, thematic links would be removed. Such links increase any topic bias effect and result in less effective and reliable measurement. The material should be relevant and if possible motivating for candidates to interact with, but the first requirement should always be good measurement.

Part 2: Group Interaction

As with Part 2 of the Writing Test, the participants expressed the view that Part 2 of the Speaking Test, the group interaction task, gave rise to the question of whether it was content knowledge or proficiency that was being tested. Further, it seemed that candidates sometimes had a set of prepared responses that enabled them to satisfactorily complete that part of the Test. According to respondents, this had led to a negative washback effect, in that instead of teachers engaging more in discussions with colleagues, they were memorising stock phrases and strategies. While it was still considered valid to have a discussion, the candidates should not be judged on metalanguage. It would be better to have different kinds of stimuli for them to discuss. There should be the possibility to move away from a discussion of errors to a discussion of different genres or text types.

The existing task focused more on content knowledge (pedagogy) than language proficiency. Similar arguments applied here as were presented for the revision of the Writing Test. Testing these features was not appropriate for a language proficiency test, and it was questionable how far the existing test format allowed a valid assessment of these skills. Correlations showed that this part of the Test was measuring a different ability from the tasks which obviously relate to language ability (0.49–0.54).

The correlation between criteria *within* Part 2 was high (0.75), but correlations with other Speaking Test scores were low.

Another problem identified with the existing test format was that it may have advantaged candidates who spoke first, leaving little for later speakers to say about the student work. The group interaction task should be more relevant to the assessment of language proficiency and should allow a fair contribution from all candidates.

The critical linguistic skill in the existing task was described in Scale 5—interaction ability. This was in contrast to Scale 6, which focused on metalinguistic and pedagogical knowledge. In this category of assessment, it was most helpful to break interaction down into more explicitly described components, for the benefit of candidates, assessors and users of the results. Interaction involves seeking and providing information, playing a cooperative and not overly dominant role, and ensuring that the task is completed.

A fundamental function of teaching is eliciting information from students, and responding to students' questions. Although the appropriateness of the way teachers interact with students was assessed in the Classroom Language Assessment (CLA), it was considered necessary to assess how well teachers can ask for and provide information, without the added complexity of the classroom situation.

Revision: A new task, involving an information gap, was to be included in the Speaking Test in place of the group interaction task.

A new task was proposed, involving an information gap, taken by test takers in groups of three (or four, where necessary). This task type was considered to have several advantages. A basic form of the task could easily be manipulated to produce numerous variations while retaining an essentially comparable task. The design should allow for the task to be completed by groups of three or groups of four test takers. Possible topics included:

- Planning an excursion, meeting a range of constraints;
- Dealing with student behaviour;
- Discussing a student's learning problems;
- Planning a school open day, fair, sports day, parent's evening, etc.
- Discussing a response to examination results; and
- Responding to enquiries from parents.

The purpose of the group interaction task would be described on the information card of each group member so that the discussion could be initiated by any of them.

Scales and Descriptors for Speaking

A revised task for Part 2 would require new scales and descriptors to be produced. The revised scales would cover the two following components:

1. *Eliciting Information.* Asking questions (including clarifying information provided by others, feedback, echoing, encouraging other to say more) and contributing sufficiently to this aspect of the interaction;

2. *Providing Information.* Answering questions clearly, relevantly, cooperatively and contributing sufficiently to this aspect of the interaction.

In order to score well on this task, candidates would need to both provide and seek information.

Revision: The scales and descriptors for Part 2 of the Speaking Test were to be revised to reflect the new task.

In addition to revising Scales 5 and 6 to take account of the new task, the other scales should be re-examined to make them easier for assessors to use. Collectively, it was hoped that these changes should make it easier for assessors to produce scores in which they had confidence, and hence improve the reliability of the paper.

Other Issues

The recommendation was made that should resources be available to implement it, and the Speaking Test should be video-recorded. This would enable more accurate rating of performances and would supply a ready-made source of sample performances for assessor training. In fact, the HKEAA revealed that plans were under way for future speaking assessments to be video-recorded. The hope was that HKEAA would trial the feasibility of doing this for LPATE with a view to the Speaking Test being video-recorded from 2008 onwards.

Other comments on the Speaking Test are summarised below:

- *For reliability, the number of assessors should be reduced. This might occur naturally due to reduced candidature.*
- *The assessment should be held at weekends to avoid tiredness.*
- *Self-access assessor training packs are suggested.*

Development and Piloting of the Speaking Test

The revision team took the above recommendations from the review into account when designing the revised Speaking Test. The requirement to read a poem was removed from the Reading Aloud section, denoted as Task 1A. In this part of the Test, candidates would now be required to read a prose passage only. The prose passages were chosen with a view to including more ‘literary’ type texts which would be more likely to be read aloud to students. In addition, the former Task 1C, now denoted as 1B, in which candidates were given a topic to talk about, was de-linked from the reading aloud task in terms of topic, so that candidates would not be doubly penalised (or advantaged) due to any unfamiliarity (or familiarity) with a single topic. The recommendation to standardise the allotted time for this task was explored. However, after careful consideration, it was decided that for logistical reasons, it would be very difficult to set a standard time for both Task 1A and Task

1B given the variations in text length and complexity as well as reading speed of the candidates. It would also require the assessors to reset the clock after Task 1A. The logistical procedures of the Speaking Test had matured thanks to the expertise of the HKEAA staff that run it and it was felt that the procedures should remain the same wherever possible. It was also felt that retaining a total time of five minutes for both the tasks in Part 1 of the Test would enable the flow of candidates to remain as is, while at the same time allowing plenty of time for the full assessment of each candidate's performance in Tasks 1A and B.

In Part 2 of the Test, the group interaction, initially four sets of tasks were designed along the lines stated above and trialed on a small group of test takers. The feedback from assessors and test takers was that by providing a certain amount of information and requiring them to extract other information from each other resulted in a situation where it was unlikely that any test taker would fail to complete the task. Moreover, it was found that test takers would read each other's task notes to obtain the information they needed. There were also concerns that by appointing roles to the candidates, there may be issues of fairness as some roles may put one or more candidate at an advantage over the others. Hence, it was decided to provide a more general scenario so that they could explore it themselves and as each would have the same information, issues of fairness would not arise. Such tasks would also be much easier to design. The 94 test takers who took part in the Pilot Tests did one (or more) of four different tasks. A total of 10 experienced LPATE Speaking Test assessors took part in the Pilot Tests. Each test taker was assessed by one assessor initially. All performances were video-recorded and then each performance was assessed a second time from the video recording by a different assessor.

The *Scales and Descriptors for Speaking* were revised to take into account both the new design of Part 2 of the Test and the concerns expressed by stakeholders during the review. The main changes can be summarised as:

- The addition of 'Lexical Accuracy and Range' to the *Grammatical Accuracy* scale to account for use of vocabulary;
- The replacement of the existing Scale 6 (*Explaining Language Matters to Peers*) to a new scale (*Discussing Teaching Matters with Peers*) to reflect the revised group interaction task;
- The change of format to a more inclusive point form for ease of use by markers.

During the assessing process the scales and descriptors were amended to take into account the opinions of the assessors.

The major change in the *Scales and Descriptors for Speaking* was in Scale 6. The focus of this scale would now be on the quality and relevance of the candidates' contributions to the discussion, rather than on their perceived knowledge of language and/or pedagogy. The task aimed to simulate the kind of discussion that teachers of English are likely to engage in practice, where they would need to contribute in staff meetings held in English. Feedback on the scales and descriptors from the assessors who took part in the Pilot Tests was very positive and they found them easier to use and clearer than the existing scales.

Table 13.17 Measure of differences in testing criteria for Speaking Test

Scale	Fair Average	Measure (logits)	Model error	Infit mean square
Pronunciation	3.0	-0.06	0.14	1.0
Reading	3.0	+0.11	0.14	0.8
Grammar	2.9	+0.38	0.14	0.7
Organisation	3.1	-0.42	0.14	1.0
Interaction	3.0	+0.05	0.14	1.2
Explanation	3.0	-0.05	0.14	1.1
Mean	3.0	0.00	0.14	1.0
S.D.	0.1	0.24	0.00	0.2

RMSE 0.14 Adj S.D. 0.19 Separation 1.37 Reliability 0.65

Fixed (all same) chi-square: 17.1 d.f.: 5 significance: 0.00

Random (normal) chi-square: 5.0 d.f.: 4 significance: 0.29

Key: Pronunciation = Pronunciation, Stress and Intonation; Reading = Reading Aloud with Meaning; Grammar = Grammatical and Lexical Accuracy and Range; Organisation = Organisation and Cohesion; Interaction = Interacting with Peers; Explanation = Explaining Teaching Matters to Peers

In order to answer the question of whether the new tasks and the revised scales and descriptors were comparable in level of difficulty as well as in the constructs of speaking that they were describing, a detailed analysis of the Pilot Test results was carried out via Many-Facet Rasch Analysis. In the case of the Speaking Test, 'item' refers to the scores on the six scales on which test takers' speaking proficiency is measured, in this case, the criteria of *Pronunciation, Stress and Intonation; Reading Aloud with Meaning; Grammatical and Lexical Accuracy and Range; Organisation and Cohesion, Interacting with Peers; and Explaining Teaching Matters to Peers*. Table 13.17 shows the analysis.

Table 13.17 shows that the criterion of *Grammatical and Lexical Accuracy and Range* proved to be the most difficult for test takers to score well on with a logit measure of 0.38. This is consistent with observations of the ways that candidates typically performed on the LPATE Speaking Test. *Organisation and Cohesion* proved to be the easiest scale to score on with a logit value of -0.42, again consistent with typical performance of candidates in the Speaking Test. The other criteria were very close in terms of measure but as the model data fit for each criterion was within the range of 0.5–1.5 logits, it was felt that these items were performing as they would be expected to.

These findings accord with those observed in other studies (see Pollitt & Hutchison, 1987; Falvey & Coniam, 2000)—where the most demanding scales tend to be those involving the formal 'expressive' categories (Pollitt & Hutchison, 1987, p. 75) of syntax, lexis, spelling.

The MFRA analysis also provided information on the different task versions used in the Pilot Tests and these are shown in Table 13.18.

Table 13.18 shows some differences in the level of difficulty of the test tasks used. This was expected as the revision team wished to gain a measure of the types

Table 13.18 Measure of differences in test tasks

Version	Task	Fair Average	Measure (logits)	Model error	Infit mean square
1	1A	3.0	-0.01	0.18	1.0
	1B	3.0	+0.01	0.18	1.1
	2	3.0	+0.00	0.18	1.2
2	1A	3.0	+0.24	0.17	0.8
	1B	3.1	-0.24	0.17	0.8
	2	3.0	+0.00	0.17	1.2
3	1A	3.0	+0.03	0.20	1.0
	1B	3.3	-0.88	0.20	0.7
	2	3.2	-0.72	0.20	1.2
4	1A	2.8	+0.71	0.33	0.7
	1B	3.0	+0.14	0.36	1.0
	2	2.8	+0.72	0.36	1.0
Mean		3.0	0.00	0.22	1.0
S.D.		0.1	0.45	0.07	0.2

RMSE 0.24 Adj S.D. 0.39 Separation 1.64 Reliability 0.73
 Fixed (all same) chi-square: 44.1 d.f.: 11 significance: 0.00
 Random (normal) chi-square: 10.4 d.f.: 10 significance: 0.40

of task that were most suitable for use in the revised Speaking Test. For Task 1A, version 4 was the most difficult reading aloud task, with a logit score of 0.71. The text chosen was the most literary of the four and involved both dialogue and idiomatic expression. For Task 1B, version 3 stood out as being the easiest task (logit value -0.88). In this task, the test takers were required to talk about learning strategies. Possibly the education focus of this task allowed the test takers to construct responses from memory based on their knowledge of the topic and the particular lexis used to describe it. For Task 2, version 3 (school outing) was the easiest and version 4 (student discipline) the most difficult. The analysis provided very useful information on the task types to be included in future Speaking Tests.

Test Result Reporting

As for the Writing Test, it was recommended that Rasch modelling be used for the Speaking Test—to both aid the process of test analysis and in the reporting of test scores and that the use of Fair Average scoring would allow a more balanced and fairer assessment of candidates' proficiency.

Summary of Revisions to Speaking Test

Based on the recommendations to emerge from the review of the LPATE, a new Speaking Test was developed (in four versions) and piloted. The results of the Tests were analysed and measures to ensure the fair reporting of candidates' performances recommended. The new Test would serve as a sample for future administrations of the LPATE and would be included in forthcoming *Guidelines for Candidates*. The revised *Scales and Descriptors for Speaking* would be included in the LPATE Specifications for use by test developers and markers, while a simplified version would be provided for candidates.

The Classroom Language Assessment

During the review process, a focus group discussion was held with eight Classroom Language Assessment (CLA) assessors. In addition, the revision team met with about twenty faculty members from the English Department of the Hong Kong Institute of Education, who used the *Scales and Descriptors for CLA* when assessing their own trainee teachers. All provided views on various aspects of this component. The main areas for discussion were training of assessors, logistical arrangements and the scales and descriptors.

Training of Assessors

Assessors felt that there should be more training and standardisation. At the time they had half-day refresher sessions, which they found very useful, but they expressed the view that a full-day of training, as is done for the Speaking Test, would give them more confidence. In addition to new assessors being given training, they felt that all assessors should be trained/retrained regularly.

Revision: More regular training for CLA assessors was to be implemented.

There was a consensus that the feedback that they received from the LPATE Team, in terms of both oral feedback and assessment statistics, was very useful and they hoped that this would continue.

Logistical Arrangements

The assessors were able to relate instances where there had been breakdowns in communication resulting in them arriving at a school to carry out an assessment, only to be told that the teacher had taken sick leave or that another assessor had been

assigned. Such instances were quite rare but nonetheless annoying. The possibility of reducing the number of visits to each teacher by one was raised, which might have gone some way to alleviating this problem. Instead of each teacher being visited twice, they could submit one video recording for assessment in addition to having one visit from an assessor. This was found to have been effective by LPATE course providers.

There were felt to be other advantages of video recording of assessments, including the saving of manpower, the archiving of performances and the potential of being able to use recordings for training purposes.

Scales and Descriptors

The assessors all referred to the scales and descriptors when assessing and felt that in general they were comprehensive. However, they felt that it might be possible to break them down into a more ‘user-friendly’ bullet-point style. This especially applied to the *Language of Interaction* scale which is very detailed.

Revision: The scales and descriptors for CLA were to be revised.

Discussion in terms of assessment procedures centred on the use of the *Scales and Descriptors for the CLA*. The points made are summarised below:

- The addition of lexical accuracy and range to the old grammatical accuracy scale so that the new domain is called *Grammatical and Lexical Accuracy and Range*;
- The revision of the pronunciation descriptors making ‘native-speaker like’ not a requirement for a high score in favour of ‘comprehensibility’;
- The expansion and emphasis in Scale 3 *The Language of Interaction* to address the need for candidates to be genuinely interacting and providing feedback to students and not just standing in the front of the class asking a series of ‘display’ type questions;
- The expansion and the emphasis in the Scale 4 *The Language of Instruction* to address the need for candidates to present and explain ‘content’ as well as to manage the process or activity in the classroom through instructions;
- The need for both Scales 3 and 4 to make an explicit statement about previously observed practices of ‘rehearsing’ the lesson before the assessor arrives and/or reading instructional language verbatim from the course book or PowerPoint slide;
- The change of format to a more inclusive point form for ease of use by assessors;
- No change in Scales 3 and 4 as it was felt that this would be too radical a change.

The following section describes the procedures that were carried out during the revision of the scales and descriptors.

Revision and Piloting of CLA Scales and Descriptors

Based on the feedback gained from the discussions with CLA assessors, a revised version of the scales and descriptors was prepared. These were presented to members of the LPATE Team and minor revisions made.

A CLA training meeting was convened with EDB part-time assessors as well as those who had been previously involved in the approved provider group carrying out CLA assessments in schools. Members of the LPATE revision team were also present at this meeting and participated in the calibration and assessing of a number of CLA video lessons [Note 10]. At this meeting the changes to the CLA descriptors were discussed and there was agreement that these should assist assessors in making the assessing easier. Three video-recorded lessons were then assessed using the revised scales and descriptors for standardisation purposes. By the end of the session there was agreement between assessors as to a fair set of scores for each of the assessed teachers. It was agreed that all those who had been standardised at this meeting should return to assess a number (10) of videoed lessons using the revised scales and descriptors. The results of the assessment were analysed using FACETS revealing that even with a small number of assessments, differences in assessor behaviour emerged. Assessors 1–3 (part-time assessors employed by the EDB) tended to be stricter in their assessing and 4–6 (assessors from one of the LPATE course providers) more lenient. In terms of criteria, *Pronunciation, Stress and Intonation* proved to be the one on which the teachers scored most highly and *Language of Instruction* the least.

Seven of the videoed lessons (A–E, H and J) had previously been assessed on the existing scales and descriptors and correlations between the scores given to them on the existing scales and those given on the revised scales are shown in Table 13.19.

Table 13.19 shows that the scores given on the seven lessons on the old scales and descriptors correlated reasonably well with the scores given on the revised scales and descriptors (though only *Pronunciation, Stress and Intonation* was statistically significant) considering the small sample size and taking into consideration the improved scale content of the revised scales and descriptors and the different ways that the constructs had been defined. Assuming the assessors of future administrations of the LPATE remained the same and they went through enhanced training and standard-

Table 13.19 Correlations between scores for lessons assessed on old and new scales and descriptors ($n = 7$)

	Grammatical and Lexical Accuracy and Range	Pronunciation, Stress and Intonation	Language of Interaction	Language of Instruction	Test Mean
Correlations	0.58 ($p = 0.169$)	0.86* ($p = 0.013$)	0.54 ($p = 0.214$)	0.43 ($p = 0.338$)	0.75 ($p = 0.053$)

*Correlation significant at 0.05 level (two-tailed)

isation procedures, the revised scales and descriptors would provide more accurate and reliable measurement of speaking ability within the classroom.

It was recommended that as for both the Writing and Speaking Tests, the scales and descriptors in their revised form should not be provided for candidates as they were designed to be used by testers and may have been open to misinterpretation. Instead, a modified and simplified version would be provided for candidates.

Test Result Reporting

As with Writing (Part 1) and Speaking, it was considered possible to report a Fair Average score for each of the criteria of CLA and an overall Fair Average score for the assessment. This would contrast with the existing process of reporting the mean scores given by each assessor on the separate scales of performance and requiring the candidate to attain a Level 3 or above on each scale in order to reach the Language Proficiency Requirement.

Summary of Revisions to Classroom Language Assessment

Based on the recommendations to emerge from the review of the LPATE, new *Scales and Descriptors for CLA* were developed and piloted. The revised *Scales and Descriptors for CLA* would be included in the LPATE Specifications for use by test developers and markers. A simplified version would be provided for candidates.

The modified and simplified versions of the revised LPATE specifications for the Writing, Speaking and Classroom Language Assessment benchmarks are provided in Appendices Ea “[LPATE Performance Descriptors: Writing \(Part 1: Composition\)](#)”, Eb “[LPATE Performance Descriptors: Speaking](#)” and Ec “[LPATE Performance Descriptors: Classroom Language Assessment](#)”.

Conclusions

This chapter has described the revision of the LPATE, in which revised tests, tasks and scales and descriptors were piloted based on the recommendations to emerge from the review of the Assessment. Revised LPATE assessment components and associated procedures and documents were produced based on the results of the Pilot Tests. Revised documents for both candidates (Candidate Guidelines) and assessment developers (Test Specifications) along with sample performances of all components were produced. The major changes made to the Assessment and endorsed by the Education Bureau are summarised below.

The major concern regarding the Reading Test was inclusion of the multiple-choice cloze passage and so this was removed and replaced by another reading passage and questions. This meant that there would be three reading passages of differing topics and genres to both ensure that only reading comprehension was being tested and candidates would have more opportunities for a fresh start (Hughes, 2003) and potential topic bias could be minimised. A similar consideration was made regarding the Listening Test, which previously had consisted of one long recording, usually an interview with one or two academics, with associated items. This was remodelled so that future tests would consist of three recordings on different topics, drawn from authentic sources and involving different speaking 'events', such as panel discussions, conversations and lectures. In addition, recommendations were made to reconsider the method of standard setting of the Reading and Listening Tests to one in which the standards of new versions of the tests would be equated to a fixed scale established using the Rasch model, rather than the existing practice of test equating with an anchor version. In fact, as will be explained in Chap. 16, the HKEAA decided to retain the method of standard setting by using the 2003 (March) papers as anchors and equating new versions of the Reading and Listening Tests to these.

Several revisions were made to the Writing Test, which would now be considered as two separate papers, Part 1 (Expository Writing) and Part 2 (Correcting and Explaining Errors/Problems in a Student's Composition). The main changes made to Part 1 were to expand the possible task types that candidates would be asked to write and to revise the scales and descriptors for rating to take account of the fact that different writing genres would be produced. Part 2, in which candidates had to identify, correct and then explain a number of perceived errors or problems in a mock student essay, was reconstructed so that error/problem correction (Task A) and explanation (Task B) were to be done on different items, thereby removing the link between the two tasks. With Part 2 now becoming an item-based test, the same standard-setting measures adopted for the Reading and Listening Tests were adopted for this part of the Writing Test.

The Speaking Test underwent quite a major revision: the somewhat controversial inclusion of reading a poem in Part 1 was dropped, leaving the reading aloud task to consist of a prose passage only. Further, the task in which candidates had to deliver a monologue on a particular topic would have its timing standardised, unlike previously in which candidates were given five minutes to deliver the poem, prose and monologue. In Part 2, where previously candidates had to discuss the perceived weaknesses and errors in a mock student composition, concerns that this was testing language knowledge rather than speaking ability were addressed by changing the task to more of a school-related one. Following the revisions, candidates were expected to engage in a meeting-like scenario in which they discuss plans for school events, policies, issues, etc. As with the Writing Test Part 1, the scales and descriptors for rating of the Speaking Test were revised. This was also the case in the Classroom Language Assessment.

The revised version of the LPATE was accepted by the Education Bureau and was launched to candidates in March 2008. In the following chapter, Neil Drave, Manager of Assessment Development for the LPATE at that time, describes how the changes were incorporated into the LPATE.

Appendix A: LPATE Revision Questionnaire



Revision of the Language Proficiency Assessment for Teachers (English Language) (LPATE)

LPATE Evaluation Questionnaire

Dear Principal/Teacher,

We have been commissioned by the EMB to conduct a revision of the Language Proficiency Assessment for Teachers (English Language) (LPATE). We would like to invite you to participate in this project because your experience and opinions are valuable input to help us carry out the revision. This questionnaire will take you approximately 20 min to complete. The information collected will ONLY be used to inform on the proposed revisions.

If you have any inquiries about this survey, please feel free to contact us. Our email addresses and numbers are listed below. Thank you very much for your help with this project.

The LPATE Revision Project Team
Hong Kong Polytechnic University

I. Professional Information (for research purposes only)

Please tick (✓) the appropriate box(es) or respond as indicated:

Please tick (✓) the appropriate box(es) or respond as indicated:

1. Professional training: Dip Ed. Cert. Ed. B.Ed.
Other (please specify) _____
2. Teaching experience: _____ years
3. Current teaching level: Primary Secondary Other (please specify) _____
4. Current post: Panel Chair
English teacher
Native English teacher
Teacher of subject(s) other than English
No current post
Other (please specify) _____

II. Experience of taking the LPATE

5. Which papers have you taken and when? Tick (✓) the appropriate box(es).

Paper	2001	2002	2003 (Mar)	2003 (Sept)	2004 (Mar)	2004 (Sept)	2005 (Mar)	2005 (Sept)	2006
Reading	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Writing	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Listening	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Speaking	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
CLA	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

6. If you have taken any of the papers more than once, please indicate your reason(s) for doing so.

To attain the passing level (Level 3)	<input type="checkbox"/>
To attain a higher level (Level 4 or above)	<input type="checkbox"/>
To improve my results	<input type="checkbox"/>
Others (please specify):	

III. LPATE Test Components

7. Please respond to the following by ticking (✓) the appropriate boxes.

Relevance		Strongly agree	Agree	Disagree	Strongly disagree	
(a) The skills tested in this paper are relevant to what the teacher uses in his/her job.	Reading	MC cloze	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
		Comprehension	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	Writing	Expository writing	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
		Error correction	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
		Error explanation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	Listening		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	Speaking	Reading aloud a poem	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
		Reading aloud a prose	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
		Giving opinions	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
		Group interaction	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	CLA		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

You may elaborate on your responses to the above. (e.g. nature/type of task/question, topic of input text, difficulty of the task, etc.)

8. Please respond to the following by ticking (✓) the appropriate boxes.

Reflection of skills			Strongly agree	Agree	Disagree	Strongly disagree
(b) The results from this paper reflect accurately a teacher candidate's ability.	Reading	MC cloze	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
		Comprehension	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	Writing	Expository writing	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
		Error correction	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
		Error explanation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	Listening		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	Speaking	Reading aloud a poem	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
		Reading aloud a prose	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
		Giving opinions	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
		Group interaction	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	CLA		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

You may elaborate on your responses to the above. (e.g. nature/type of task/question, topic of input text, difficulty of the task, etc.)

IV. The LPATE Syllabus Specifications and Guidance Notes for Candidates

9. Please respond to the following by ticking (✓) the appropriate boxes.

	Strongly agree	Agree	Disagree	Strongly disagree
(a) The Syllabus and Guidance Notes helped me to prepare for the tests.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
(b) The scales and descriptors for Writing, Speaking and CLA helped me to understand the requirements for these papers.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
(c) The Syllabus and Guidance Notes should be revised.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

You may elaborate on your responses to the above items (a), (b) and (c) below.

V. LPATE Administration

10. Please indicate how satisfactory you found the LPATE administration procedures by ticking (✓) the appropriate boxes.

	Very satisfactory	Satisfactory	Unsatisfactory	Very unsatisfactory
(a) Registration	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
(b) Briefing seminar	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
(c) Test arrangements	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
(d) Announcement of results	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

You may elaborate on your responses to the above items (a), (b), (c) and (d) below.

VI. Additional information

11. Do you have any other opinions on the LPATE that you would like to express?

Thank you for taking the time to do this questionnaire.

Appendix B: Item Intents for the Revised Reading Test

Item Intent	No. of items
Locating/identifying:	
Information making a simple inference	3
The referent of a cohesive device	2
A synonym for an idiomatic or uncommon expression using contextual cues	2
Both parts of a comparison in a specified paragraph in the presence of conflicting information	1
A synonym for a linking phrase in context	1
A synonymous match in the presence of competing information	1
The referent of a cohesive device which links ideas across paragraphs	1
The meaning of an uncommon word from the context	1
The altered meaning of a familiar word in context, making connections across a paragraph	1
The reason for a specified phenomenon in a paragraph	1
Explicitly stated information given in adjacent sentences	1
Suggestions for action made by an author, embedded within text	1
Agreement or inconsistency/opposition between each of a series of statements and arguments put forward in a text	1
Lexical clues to identify relevant section of text and interpret potential effects of actions	1
An aspect of an author's complex attitude on an issue	1
Interpreting:	
An idiomatic expression using contextual cues	6
Information in a specified paragraph	3
Information in a specified paragraph in the presence of competing information	1
The intent or effect of a metaphor	2
The meaning of two metaphors	1
The meaning of an unknown word using contextual clues	1
The underlying meaning/implications of a familiar word used metaphorically in a specific context	1
A colloquial expression using tone and other contextual clues	1
An attitude implied in a specified paragraph	1
Atypical use of uncommon lexical expression in context	1
The perspective taken in a specified part of the text	1
Reason for and effect of including comparison to body outside text	1
Understanding:	

(continued)

(continued)

Item Intent	No. of items
The main point of a specified paragraph which contains idiomatic expressions	1
The meaning of a common word used idiomatically, and demonstrating this understanding by finding a synonym in the text	1
The underlying criticism made in a specified paragraph	1
The main message of a specified passage by interpreting idiomatic language through contextual clues	1
The meaning of a linking phrase in context	1
The effect and purpose of a pun	1
Irony embedded in a paragraph	1
Connections across a paragraph to understand point of reference	1
Reflecting:	
On and evaluating overall content of an article, giving own opinion about suitability of title	1
On intent of fairly common textual convention (ellipsis, repetition)	1
Extracting:	
Two main messages from a grammatically challenging sentence, using own words to explain these main messages	1
Explaining:	
The likely reason for an image included with the text by recognising its relevance to the subject matter	1

Appendix C: Item Intents for the Revised Listening Test

Item Intent	No. of items
Locating/identifying	
Specific information	9
Information about the speakers	4
Information in the presence of competing information	4
Two metaphors mentioned by speaker	2
Lexical use by the speaker	1
Interpreting:	
The reason for a specified phenomenon given by the speaker	8
The perspective taken on a specified issue	5
The speaker's point of view	3
The intent or effect of a metaphor	2
The meaning of an unknown word using contextual clues	1
An idiomatic expression using contextual cues	1
The reasoning of the speaker	1
The intention of the speaker	1
Understanding:	
The speaker's point of view	3
An illustration intending to make a point	2
Summarising:	
Main ideas from a section of text	6
Extracting:	
Required information from conflicting information	3
Explaining:	
Actions reportedly taken by the speaker in certain circumstances	1

Appendix D: Item Intents for the Revised Writing (Part 2) Test

Item Intent	Item No.
Task A	
Incorrect prepositional phrase	2
Lack of plurality in noun	2
Incorrect preposition	2
Incorrect formation of past tense	3
Use of incorrect participle	3
Incorrect use of comparative adjective	4
Incorrect verb formation	4
Incorrect tense	4
Incorrect verb choice	5
Use of unnecessary preposition	5
Incorrect word order	5
Incorrect formation of pronoun	6
Incorrect choice of lexical item	6
Use of incorrect part of speech	6
Unnecessary pluralisation of gerund	7
Use of run-on sentence	7
Omission of definite article	7
Use of incorrect participle to form past continuous tense	8
Omission of verb	9
Use of incorrect verb	10
Task B	
Incorrect use of infinitive	11
Incorrect use of adverb	12
Lack of subject–verb agreement	12
Omission of relative pronoun to introduce relative clause	13
Use of infinitive instead of participle in forming past continuous tense	14
Incorrect formation of past perfect tense	15
Use of run-on sentence	15
Use of inappropriate conditional sentence through incorrect conjunction	16
Use of incorrect prepositional phrase	17
Incorrect choice of verb	18

Appendix Ea

LPATE Performance Descriptors: Writing (Part 1: Composition)

(Note: The descriptors are for illustrative purposes to help candidates to grasp the skills required at each level. They are a simplified version of the scales and descriptors used by assessors in the assessment of performance in the LPATE.)

For Writing (Composition), candidates are assessed on the following three scales:

- Organisation and Coherence (OC)
- Grammatical and Lexical Accuracy and Range (GLAR)
- Task Completion (TC)

The following descriptors indicate what candidates are expected to be able to do at each level on this task.

5	OC	Writes a completely coherent text such that ideas and information flow in a smooth and natural way. Makes use of appropriate language to ensure cohesion and logical links between ideas
	GLAR	Demonstrates control over a range of grammatical structures and vocabulary, including idiomatic expressions
	TC	Addresses all elements of the task, with elaboration and illustration where appropriate
4	OC	Writes a coherent text such that ideas and information flow in a mostly smooth and natural way. Makes use of appropriate language to aid cohesion and logical links between ideas
	GLAR	Demonstrates control over a range of grammatical structures and vocabulary, including idiomatic expressions, though with occasional mistakes
	TC	Addresses all elements of the task, with some elaboration and illustration
3	OC	Presents ideas and information in a generally clear way. Links ideas together using mostly appropriate language
	GLAR	Demonstrates a limited control over grammatical structures and vocabulary
	TC	Completes the task with minor omissions
2	OC	Presents ideas and information in a way that makes it difficult for a reader to follow. Does not link ideas effectively
	GLAR	Demonstrates a very limited control over grammatical structures and vocabulary
	TC	Fails to address one or more major requirements of the task
1	OC	Presents and links ideas and information in a way that is very difficult to understand
	GLAR	Demonstrates no control over grammatical structures and vocabulary
	TC	Does not complete the task

Appendix Eb

LPATE Performance Descriptors: Speaking

(Note: The descriptors are for illustrative purposes to help candidates to grasp the skills required at each level. They are a simplified version of the scales and descriptors used by assessors in the assessment of performance in the LPATE.)

For Speaking, candidates are assessed on the following six scales:

Part 1 Task 1A:	Pronunciation, Stress and Intonation (PSI) Reading Aloud with Meaning (RAM)
Part 1 Task 1B:	Grammatical and Lexical Accuracy and Range (GLAR) Organisation and Cohesion (OC)
Part 2:	Interacting with Peers (IP) Discussing Educational Matters with Peers (DEMP)

Please refer to the DVD entitled Language Proficiency Assessment for Teachers (English Language): Speaking and Classroom Language Demonstration for sample performances.

The following descriptors indicate what candidates are expected to be able to do at each level on each task.

Task 1A: Reading Aloud

5	PSI	Reads in a fully comprehensible way with no systematic errors in pronunciation and uses stress and intonation in a very natural way
	RAM	Uses speed and pausing in a very natural way to convey the meaning of the text.
4	PSI	Reads in a comprehensible way with few systematic errors in pronunciation and uses stress and intonation in a mostly natural way
	RAM	Uses speed and pausing in a mostly natural way to convey the meaning of the text
3	PSI	Reads in a generally comprehensible way, though may make errors in pronunciation. Uses stress and intonation to convey meaning, though may occasionally sound unnatural
	RAM	Uses speed and pausing to convey the meaning of the text, despite sounding occasionally unnatural or inappropriate
2	PSI	Does not read in a consistently comprehensible way due to errors in pronunciation, stress and intonation and speech is frequently hesitant
	RAM	Does not convey the meaning of the text effectively through the use of speed and pausing. May be monotonous or overly dramatic
1	PSI	Makes frequent errors in pronunciation, stress and intonation which cause confusion for the listener
	RAM	Speed and/or pausing are not used in any consistent way. Does not convey the meaning of the text

Appendix Ec

LPATE Performance Descriptors: Classroom Language Assessment

For Classroom Language Assessment, candidates are assessed on the following four scales:

- Grammatical and Lexical Accuracy and Range (GLAR)
- Pronunciation, Stress and Intonation (PSI) ;
- The Language of Interaction (L-Int);
- The Language of Instruction (L-Instr).

The following descriptors indicate what candidates are expected to be able to do at each level.

5	GLAR	Always able to use an appropriate range of grammatical structures and vocabulary accurately
	PSI	Speaks in a fully comprehensible way with no systematic errors in pronunciation and uses stress and intonation in a very natural way to convey meaning
	L-Int	Maintains very smooth interaction with students using a range of effective and appropriate language
	L-Instr	Presents and explains lesson content clearly and naturally and provides clear instructions
4	GLAR	Uses an appropriate range of grammatical structures and vocabulary mostly accurately
	PSI	Speaks in a comprehensible way with few systematic errors in pronunciation and uses stress and intonation in a mostly natural way to convey meaning
	L-Int	Usually maintains smooth interaction with students using a range of effective and appropriate language
	L-Instr	Usually presents and explains lesson content clearly and naturally and provides clear instructions
3	GLAR	Uses a range of grammatical structures and vocabulary generally accurately, though with occasional errors
	PSI	Speaks in a generally comprehensible way, though may make errors in pronunciation. Uses stress and intonation to convey meaning, though may occasionally sound unnatural
	L-Int	Generally able to interact with students using appropriate language
	L-Instr	Usually presents and explains lesson content and provides instructions effectively, though may at times sound repetitive and unnatural
2	GLAR	The range of grammatical structures used is limited and consistently inaccurate. Vocabulary is limited
	PSI	Does not speak in a consistently comprehensible way due to errors in pronunciation, stress and intonation and speech is frequently hesitant

(continued)

(continued)

	L-Int	Does not interact with students effectively due to limited appropriate language
	L-Instr	Often does not present or explain lesson content or provide instructions effectively
1	GLAR	A very limited range of grammatical structures and vocabulary is used. Fails to convey meaning due to frequent grammatical errors
	PSI	Makes frequent errors in pronunciation, stress and intonation which cause confusion
	L-Int	Does not interact with students due to a lack of appropriate language.
	L-Instr	Fails to present or explain lesson content or provide instructions effectively

Notes

1. The LPAT Main Committee was the overseeing body responsible for the running of the LPAT (English Language) and LPAT (Putonghua). It consisted of representatives from Government (Education Bureau), the Hong Kong Examinations and Assessment Authority, tertiary institutions (including universities and teacher education providers), school principals, school teachers and employers from the public and private sectors.
2. There were single administrations per year in 2001, 2002 and 2006 and two per year in 2003, 2004 and 2005. The final administration of the LPATE in its existing form took place in 2007.
3. For the purposes of comparing performances and standards between the new versions of the test components and the existing ones, the Reading, Writing and Listening Tests from March 2003 were used (with permission from the HKEAA). These tests were administered to a cohort of approximately 2000 candidates. The test papers have remained secure since then as publication of the LPATE question papers did not commence until after the September 2003 administration.
4. Studies of performance on multiple-choice cloze exercises in language tests have shown that these types of exercise do not provide evidence of being able to measure distinct skills or abilities. (For studies of the effectiveness of multiple-choice cloze exercises, see Hale et al., 1988; Abraham & Chapelle, 1992).
5. Currently, during the Expert Judgment standard-setting exercise, the judges make decisions based on the evidence of candidates' performance by looking at live scripts. For the multiple-choice cloze, they were unable to do this and had to take as a reference point the mean score of all candidates on the multiple-choice cloze part of the paper. This potentially compromised the integrity of the process.
6. Another possible item type was the C-test, where the first letter or two of the required word is provided. This would, however, have required a considerable amount of investigation to see how well this kind of item worked in the Hong Kong context.

7. Radio Television Hong Kong (RTHK) is the public broadcasting service of Hong Kong.
8. For Fair Average scores, which are reported to one decimal place, a Level 5 was awarded to those attaining a score of 4.8 and above; a Level 4.5 to those scoring 4.3–4.7; a Level 4 to those scoring 3.8–4.2, and so on.
9. No reliability figures are available for previous LPATE Writing (Part 2) Tests. For the 2003 (March) test used in the Pilot Tests, reliability figures of 0.46 for Task A and 0.56 for Task B were returned.
10. The cut scores for Writing (Part 2) of the LPATE were fixed, with candidates being required to get 60% of items correct to attain Level 3 on each of the two tasks. This equated to a score of 12 out of 20.
11. The video performances were supplied by the Centre for Professional and Business English of the Hong Kong Polytechnic University.

References

- Abraham, R. G., & Chapelle, C. A. (1992). The meaning of cloze test scores: An item difficulty perspective. *The Modern Language Journal*, 76(4), 468–479.
- Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.
- Barkaoui, K. (2010). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly*, 44(1), 31–57.
- Blumberg, H. H., Fuller, C., & Hare, A. P. (1974). Response rates in postal surveys. *The Public Opinion Quarterly*, 38(1), 113–123.
- Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.
- Coniam, D. (2008a). An investigation into the effect of raw scores in determining grades in a public examination of writing. *Japan Association for Language Teaching Journal*, 30(1), 69–84.
- Coniam, D. (2008b). Problems affecting the use of raw scores: A comparison of raw scores and FACETS' fair average scores. In L. Taylor & C. J. Weir (Eds.), *Multilingualism and assessment* (pp. 179–190). Cambridge: Cambridge University Press.
- Coniam, D., & Falvey, P. (2001). Awarding passes in the language proficiency assessment of English language teachers: Different methods, varying outcomes. *Education Journal*, 29(2), 23–35.
- Denscombe, M. (1998). *The good research guide for small-scale social research projects*. Buckingham, UK: Open University Press.
- Education Bureau, Hong Kong. (2005). *Revision of the Language Proficiency Assessment for Teachers (LPAT) (English Language/ Putonghua)*. Hong Kong: Government Printer. Unpublished document.
- Eckes, T. (2009). Many-Facet Rasch measurement. In Takala, S. (Ed.), *Reference Supplement to the manual for relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment (Section H)*. Strasbourg, France: Council of Europe/Language Policy Division, http://www.coe.int/t/dg4/linguistic/manuel1_EN.asp?
- Falvey, P., & Coniam, D. (2000). Establishing English language writing benchmarks for primary and secondary teachers of English language in Hong Kong. *Hong Kong Journal of Applied Linguistics*, 5(1), 128–159.
- Filipi, A. (2012). Do questions written in the target language make foreign language listening comprehension tests more difficult? *Language Testing*, 29(4), 511–532.
- Geranpayeh, A., & Taylor, L. (2008). Examining listening: Developments and issues in assessing second language listening. Cambridge ESOL: Research Notes, 32, pp. 2–5. Available at: www.tpcbridgeesol.org/rs_notes/rs_nts32.pdf.

- Government of the Hong Kong Special Administrative Region. (2000). *Syllabus specifications for the language proficiency assessment for teachers (English Language)*. Hong Kong: Government Printer.
- Gronlund, N. E. (1985). *Measurement and evaluation in teaching*. New York: Macmillan.
- Hale, G., Stansfield, C. W., Rock, D. A., Hicks, M. M., Oller, J. W., & Butler, A. A. (1988). Multiple-choice cloze items and the TOEFL test. *Language Testing*, 6(1), 47–76.
- Hughes, A. (2003). *Testing for language teachers* (2nd ed.). New York: Cambridge University Press.
- Pollitt, A., & Hutchinson, C. (1987). Calibrated graded assessment: Rasch partial credit analysis of performance in writing. *Language Testing*, 4, 72–92.

Alan Urmston is an Assistant Professor in the English Language Centre at the Hong Kong Polytechnic University. As well as teaching on undergraduate and postgraduate programmes, he coordinates assessments for the Centre. His main publication and research interests are in language assessment and sociolinguistics.

Chapter 14

Maintaining Standards in the Indirectly Assessed Components of the LPATE



Neil Drave

Abstract This chapter discusses how standards are maintained in the indirect (analytically marked) components of the LPATE. It should be remembered that two of the five LPATE components are wholly direct (performance) tests of teacher language skills, as is Part I of the Writing component. The remaining two components of the LPATE, the Reading and Listening Tests, plus the parts of the Writing Test which concern errors, are marked analytically. It is important to reconcile the two modes of marking. This chapter shows how cut scores are created for the analytically marked components of the LPATE and describes the standards-setting mechanism for the English language LPATE. The LPATP (for Putonghua) employs a different standards-setting mechanism. An earlier version of this chapter was presented at the 36th International Association for Educational Assessment (IAEA) Conference, Bangkok, Thailand (2010), and was distributed electronically as ‘Keeping up appearances: maintaining standards in Hong Kong’s LPATE’. The PDF is available at <http://www.iaea.info/papers.aspx?id=78>.

Introduction

Since its introduction in 2001, the LPATE has been a high profile, and often controversial, part of the HKSAR Government’s strategy for raising English teaching standards and making the profession more ‘professional’ (Qian, 2008; Drave, 2006). LPATE was the first standards-referenced assessment offered in Hong Kong and paved the way for the adoption of standards-referenced reporting (Great Schools Partnership, 2017) in the Hong Kong Diploma of Secondary Education (HKDSE), the suite of public examinations taken by the majority of school leavers, which was implemented in 2012.

N. Drave (✉)

Hong Kong Examinations and Assessment Authority, 12/F, Southorn Centre,
130 Hennessy Road, Wan Chai, Hong Kong
e-mail: ndrave@hkeaa.edu.hk

© Springer Nature Singapore Pte Ltd. 2018
D. Coniam and P. Falvey (eds.), *High-Stakes Testing*,
https://doi.org/10.1007/978-981-10-6358-9_14

311

The gate-keeping function of the LPATE, its high profile within the educational community and its pioneering reporting method make the LPATE an important and influential assessment. It is therefore crucial that the standards which its candidates must meet are well thought out, clearly articulated and reliably implemented from year to year.

In this chapter, the current LPATE standard setting and maintenance mechanisms are described, focusing on the setting of cut scores in the indirectly assessed papers, and then evaluated in light of current research. The chapter concludes with a list of issues which need to be considered when setting standards. These issues have already been touched on in Section I, Chap. 8.

Direct and Indirect Components of the LPATE

Of the five LPATE components, the Writing Test, Part 1 (Composition), the Speaking Test and CLA are direct performance assessments which employ explicitly defined scales and descriptors. The method of setting the standard in these 'productive' skills is quite straightforward: samples of performance from previous years (e.g. candidate compositions, videos of Speaking Tests) are reviewed by senior examiners, who decide what level they demonstrate, and these are then shown to new markers in standardisation and training sessions. The aim is to ensure that the observed performance is matched with the descriptors at each scale level and that the required standard does not change from year to year.

The other LPATE papers (Reading, Writing [detection and correction of errors/problems] and Listening) are pen-and-paper tests which give rise to numerical scores. The original proficiency levels for the papers were defined by a panel of experts using data from the PBAE, administered in 1999 (Coniam & Falvey, 1999). There were revisions to the format of the LPATE in 2007, and these necessitated a re-setting of the standards, which was undertaken using the same procedures as are currently used for the LPATE (see below). It is therefore the equivalents of these levels which need to be identified for the tests in each administration to ensure that the prevailing standards match those which are publicly available.

LPATE employs a 'conjunctive' rather than a 'compensatory' scoring method (Zieky & Perie, 2006, p. 10; Hambleton & Pitoniak, 2006, p. 450), meaning that a candidate has to pass all scales in all papers to be benchmarked: there is no provision for failure on one scale to be compensated for by a pass on another. Conjunctive methods are considered appropriate for tests of separate language skills (Zieky & Perie, 2006, p. 10) since it is fair to expect a candidate to be competent in all areas and because the LPAT is a high-stakes test of teacher language ability. The LPATE employs within-paper conjunctive scoring, a practice which may lead to lower pass rates than compensatory methods (Coniam & Falvey, 2001), which has to be taken into consideration when setting standards.

Standard Setting (Maintenance) Methods

The task of the ‘standards-maintaining’ exercise (Bramley & Black, 2008, p. 3) for the LPATE paper-and-pencil tests is to identify points on a score scale that divide the observed test score into classifications on a five-point scale. Decisions have to be made as to what score ranges are considered equivalent to LPATE Levels 1, 2, 3, 4 and 5, i.e. what the cut scores should be (Kane, 2001, p. 55) in order to ensure that the 2008 standards are maintained in subsequent years. There is no foolproof way of making these classification decisions, and all methods have their strengths and weaknesses (Cizek, 2001; Hambleton & Pitoniak, 2006). For example, any score is subject to measurement error, including cut scores. There are also political and other pressures on test administrators, meaning that the process of determining cut scores may be as much political as it is educational (Zieky, 2001, p. 46; Kane, 2001, p. 58; Cizek, Bunch, & Koons, 2004, p. 32). Nevertheless—or perhaps because of this—setting cut scores is a crucial task.

It is common to categorise standard setting methods into those based on judgements about test questions and of test takers’ work (Hambleton & Pitoniak, 2006; Zieky & Perie, 2006; Cizek et al., 2004). Test questions are the focus of the Nedelsky Method for multiple-choice items, the *Bookmark Method*, in which items are ordered according to difficulty, and the *Ebel Method*, which requires judgements about item relevance and importance as well as difficulty. Candidate work is monitored in the *Borderline and Contrasting Groups Methods*, as well as in the *Body of Work Method*, in which holistic ratings are given to ‘response booklets’ containing a number of test-taker performances. LPATE draws upon both categories of methods (item- and candidate-focused) to maintain its standards.

LPATE Practice

The two methods of determining cut scores in the LPATE are Expert Judgement and Rasch Analysis. The Expert Judgement Method (which might more properly be termed a ‘procedure’ since it can be used with a number of methods) is a variation of the Extended (or Modified) Angoff Method (Hambleton & Pitoniak, 2006; Zieky & Perie, 2006), used for many purposes, including linking language tests to the Common European Framework of Reference for Languages (CEFR) (Tannenbaum & Wylie, 2009), the certification of doctors in the USA (Morrison, McNally, Wylie, McFaul, & Thompson, 2009; Clauser, Mee, Baldwin, Margolis, & Dillon, 2009), and of medical translators in Japan (Kozaki, 2004). In this method, a panel of expert judges is asked to estimate the proportion of target examinees that would probably be able to answer correctly each of the test items. For the LPATE, the target examinees are minimally competent teachers of English in Hong Kong schools. The proportion assigned to each item may also be conceptualised as the probability that a single target examinee can answer a particular item correctly. The judges’ estimations are

expressed in the form of a numerical value between zero and one: '0.65' would mean a 65% probability of a correct response for example. A cut score is then obtained by summing these values across all items. The Angoff Method originally required simple 'Yes' or 'No' responses and was tailored to MC items, but the modification of requiring percentages has proven to be very suitable for constructed-response items (Cizek, 1993, 1996; Berk, 1986, 1995; Hurtz & Auerbach, 2003).

This extended Angoff Method is supplemented in two ways in the LPATE: first, by providing descriptive test statistics and other relevant data (e.g. normative or impact figures); and second, by giving judges an opportunity to directly review live candidate scripts (cf. Hambleton & Pitoniak, 2006, p. 447).

Before the judges make their final decision on what cut scores to recommend, they are given additional data in the form of cut scores suggested by a one-parameter IRT (Rasch) analysis. Rasch has many uses in the social sciences (Bond & Fox, 2001) and is widely used for setting educational standards (e.g. in pharmaceutical examinations, Jackson, Draugalis, Slack, Zachry, & D'Agostino, 2002). Rasch analysis is described more fully in Appendix A 'Methodological Approaches and Analytical Tools' in Chap. 8 at the end of Section I. The LPATE test-equating data come from common person equating of pretests and anchor tests, the results of which are extrapolated to the live versions of the tests. Rasch can be used to provide data which might help judges decide on their probabilities, such as item difficulty figures (Bond, 2003, p. 191; MacCann & Stanley, 2006; Tiratira, 2009), but in the LPATE it is used only for test equating.

The current LPATE **maintenance procedure** may be summarised as follows:

1. **Pretesting**

An anchor test is administered for each LPATE indirect paper in tandem with a trial version of the live test (with more items). The anchor is publicly available and the cut scores for each level clearly stated (Education Bureau, 2010). Two or more versions of the pretests are prepared to ensure that data are obtained for all sections of the test and that there is less chance of adjacently seated pretest candidates copying from each other.

2. **Test equating**

After marking the pretest, and the deletion of unsuitable items, common person test equating is undertaken. Using the RUMM computer program, the pretest candidates' scores on both tests are converted to logits (Bond & Fox, 2001; Luo, Seow, & Chew, 2001; Yu & Popp, 2005). The logit values which correspond to the published (anchor) Levels 2, 3, 4 and 5 cut scores are extrapolated to the live tests. First, the logit values which correspond to the appropriate cut scores on the anchor test are identified. Second, the same logit values on the live test are identified. Third, the live test scores which correspond to these logit levels are read off and recorded for later use.

3. **Statistical analysis**

When the live test has been administered and marked, test statistics are calculated. The most useful for standards-maintenance purposes are the measures of central tendency, the IF index (item facility, i.e. percentage correct) and the ID index (item discrimination, i.e. how well each item was answered by the best and worst candidates).

4. **Expert Judgement meeting**

An Expert Judgement exercise (in the form of a round-table meeting) is then carried out to:

1. Gather information from a source other than the Rasch analysis to help determine the cut scores.
2. Ensure that the determination of the cut scores for the three papers is transparent to, and monitored by, members of different stakeholder groups. This is important because it ensures that the process has face validity.

The expert group normally has between 10 and 15 participants, which is an optimal number (Zieky & Perie, 2006). It includes primary and secondary school principals and teachers, lecturers from tertiary institutions, representatives from committees which designed and moderated the test papers and representatives from the HKSAR Government's Education Bureau, which originally commissioned the LPATE.

Judges are sent an information pack containing the test papers and instructions and asked to work through the papers as if they were candidates before they attend the Expert Judgement exercise. Each panel member then estimates the probability of a just-qualified teacher answering each item correctly, as described above. The estimates are entered into a spreadsheet, outliers are removed, and the probabilities summed to give preliminary cut scores.

The expert panel begins by setting the cut score for Level 3 on one of the papers (normally the Reading), as Level 3 is the crucial 'benchmark' level.

The panel is first presented with the probability scores of all members and the preliminary cut score for Level 3 derived from the Angoff Method. The cut scores of individual panel members are compared with the cut scores suggested by the Rasch analysis. If the results of the panel are reasonably coherent and deemed close enough to the Rasch scores, the process ends and the panel makes a firm recommendation to the approving body. If the results of the panel are reasonably coherent but not close to the Rasch cut score, there are two options:

1. The panel provides an explanation to support its recommendations; or
2. The panel may revise their item probabilities.

If the second of these options is chosen, additional information is provided to the panel to help them decide whether to change their probabilities (e.g. test statistics from the live administration, impact data on how different cut scores would affect the passing percentage). A new cut score is calculated, after removing outliers,

and presented to the panel for their consideration. If there is still a difference between this figure and the one suggested by the Rasch procedure, or if there is little difference but the panel so wishes, samples of performance awarded different raw scores may be viewed. Judges may be directed to consider candidate performances on items at certain levels of difficulty (as defined by different item facility scores). A cut score is then decided upon, taking into account a one or two mark margin of error (depending on the paper), which is incorporated to account for the fact that a cut score has a standard error (Jaeger & Mills, 2001, p. 329). This procedure is then repeated for all the levels on all the papers.

Certain features of the procedure used in LPATE are known to lead to greater inter-expert consistency. Such features include discussion among judges, the opportunity to re-rate and the availability of item statistics and impact data. Giving judges access to test data leads to more realistic judgements as they then know how all the candidates, not just minimally competent ones, have actually performed (Zieky & Perie, 2006, p. 10). It also ensures that there is less of a mismatch between expert judgements of what candidates can do (which may be more or less prescriptive, as discussed below) and what they actually can do. In some contexts, however, these features may be considered as ‘contaminating the procedure’, (Zieky & Perie, 2006; Zieky, 2001, p. 37), and they may give rise to standards which are more conservative than they might otherwise be (Cizek, 2001, p. 11).

Strengths of the Current Method

The main strengths of the current LPATE standards-maintenance procedure are as follows:

1. The method has been used for many years, and all those involved know the process thoroughly. It therefore requires minimal (re)training to use.
2. The expert group is relatively stable, with a low turnover of members from year to year, and has representatives from the major stakeholder groups.
3. The group members have relevant subject knowledge and professional expertise, so they are able to make informed judgements (Cizek et al., 2004, p. 34).
4. The Extended Angoff Method leads to a single, clear-cut score for each level, which provides a firm foundation for subsequent discussion.
5. The process has objective (i.e. statistical) and subjective components, which provide a check on each other.
6. The process involves consideration of pretest and live test information, as well as impact data, which gives the judges all they need to make principled and realistic decisions about cut scores.
7. The process—at least at the meeting stage—is efficient because much of the time-consuming work (e.g. assigning probabilities, test equating) is done before the actual standard setting process gets underway.

8. The entire process conforms to (most of) the commonly accepted guidelines for good standard setting practice (see Hambleton & Pitoniak, 2006 for a list).

Issues

While the above procedure is generally satisfactory, and usually runs smoothly, there are some issues to consider when implementing it to ensure the best possible outcomes.

1. Difficulty of Judging Probabilities

- a. The Angoff Method is not easy to implement as it relies on experts' ability to internalise and apply standards from year to year as they operationalise the concepts of 'benchmark level' and 'just-qualified candidate' (Zieky, 2001, pp. 35–37; Jaeger & Mills, 2001, p. 335; Bramley & Black, 2008, p. 9). Judges' estimations of difficulty may be inaccurate. There is some research evidence, for example, that easier items are systematically judged to be more difficult than they really were, and vice versa (Hambleton & Pitoniak, 2006). This problem cannot be completely solved but is certainly mitigated by the use in LPATE of different sources of evidence about candidate performance.
- b. Judges may not be clear about their task as they are not normally asked to make probability judgements. A useful indication of whether the judges understand what they have to do is to look at the probabilities assigned to MC items, which for a four-option item should be ≥ 0.25 i.e. 1 divided by the number of choices (Zieky & Perie, 2006, p. 14). Even if a candidate were to guess the answer to an MC item, they would still have a one in four chance of getting it right, so logically the judge's probability cannot be less than one in four, or 25%.

2. Gate Keeping

Since the LPATE is a test for teachers, with teachers and teacher trainers on the expert panel, some members may be prescriptive rather than descriptive when making judgements; i.e., they may assign probabilities according to what they think candidates should be able to do rather than what they are probably capable of. Prescriptivism can be used to set cut scores, but it leads to high standards, as the experts exercise their perceived gate-keeping function (Hambleton & Pitoniak, 2006, p. 442). It is important, therefore, to discuss this issue openly and to caution against excessive strictness when briefing the expert panel members.

3. Groupthink

Like any group, an expert judgement group is subject to 'groupthink' (Janis, 1972) which manifests itself as behaviour consistent with a belief that any decision made by a group is inherently and necessarily better than that made in another way.

This may lead to a rejection of the Rasch evidence, for example, or to the stifling of dissenting voices in the group discussion phase. To counteract this tendency, care must be taken to respect the opinions of all judges, even those with dissenting opinions.

4. Pretesting and Ras(c)h Decisions

The cut scores provided by Rasch analysis are useful only in so far as the pretest and test equating data are reliable. While the HKEAA has confidence that they are, constant vigilance is required in this regard and care must be taken to ensure the following:

- (a) The pretest candidate population is large and heterogeneous enough to enable accurate calibration of items (Hambleton & Jones, 1993).
- (b) Candidates are motivated to answer to the best of their ability.
- (c) Each test item discriminates equally (Henning, Hudson, & Turner, 1985).
- (d) Items are pitched at an appropriate level of difficulty for the pretest candidates, so that item difficulty can be reliably estimated (Bond & Fox 2001, p. 58).
- (e) The marking of the anchor test is consistent from year to year.
- (f) The live test is marked in the same way as the pretest.

Additionally, any lack of understanding of the Rasch model could lead to inappropriate interpretation of its results or to judges simply ignoring them.

5. Judging Candidate Proficiency

The final stage of the expert judgement process is often the viewing of candidate work, particularly when setting cut scores for levels other than Level 3. There are various matters to be taken into account when engaging in this practice.

- (a) Because of the relatively small candidature, there are only a small number of scripts at particular mark levels, particularly at the top and bottom of the ability range. It may therefore be difficult to get a sense of the quality of candidate performance at each level.
- (b) LPATE is not wholly a criterion-referenced test, and therefore, it is difficult to give a definite point of mastery/non-mastery which would correspond to a 'benchmark' level.
- (c) Descriptors are widely considered to be a useful tool for guiding the viewing of candidate scripts/performances and making judgements about them (Cizek et al., 2004, p. 34; Alderson, 1995, p. 76; Zikey & Perie, 2006; Kane, 2001, p. 56). The HKEAA uses these for the LPATE components which assess productive proficiency (Writing Part 1, Speaking, CLA) but not for other components because it is felt that any set of descriptors would be too general to be useful for analysing a candidate's performance on an item-based test. Care must be taken, therefore, to ensure that judgements are made by comparing candidate performance to the 'putative standard'

which has emerged from previous discussion, rather than simply to other candidates, which would contravene the spirit of the standards-referenced assessment method.

- (d) Due to time constraints, there may be a tendency to look only at certain key items, which may then become the de facto criterial items for benchmarking. This tendency may be exacerbated by the fact that scripts are presented to the meeting on a computer screen rather than on paper in the LPATE meeting as this means that judges' attention at any one time is focused on a specific part of a script.
- (e) Judges must be made familiar with the criteria used to judge candidate answers so that they do not make judgements based on irrelevant criteria (e.g. handwriting).

6. Leader Integrity

The person who leads the standard setting process provides the data and leads the discussion is in a very powerful position and can influence the ultimate outcome of the process. It is therefore a matter of the utmost importance that this person is reliable. Usually, the Assessment Manager from the HKEAA leads the setting process. It is vital that they are independent and do not have a vested interest in seeing a certain proportion of candidates pass and that they keep a record of the process which can be scrutinised. Many of the issues mentioned above can be managed if the standard setting process is handled with integrity and professionalism.

A further issue concerns consistency across papers. The use of the Angoff Method involves summing the scores on items to arrive at a cut score and therefore produces a standard which is compensatory (Hambleton & Pitoniak, 2006, p. 450); candidates can be weak in one of the three passages used to test Reading, for example, but as long as they are strong enough in the other passages they can pass, as there is just one cut score. However, the direct tests of Writing and Speaking are not compensatory (or are so in a very limited sense). Different papers are therefore treated in different ways. There is a need for more research on the consequences of this and on how to ensure that the standard of English proficiency required by different papers is similar.

Conclusion

In this chapter, the process of maintaining standards for the three LPATE papers which use cut scores has been described. Given the high-stakes nature of the assessment, it is important that the method adopted is undertaken with great transparency and rigour. There is no perfect procedure, of course, and the one currently used is not without its issues. The HKEAA is confident, however, that the current practice is theoretically defensible and effectively monitored and that it gives rise to standards which are quite stable from year to year.

The following chapter in this section examines the impact of the LPATE through the eyes of the media and the possible consequences of this perspective on whether LPATE has been able to promote educational change. It foreshadows and contrasts with some of the discussion in Section IV which is based on hard data rather than writers' opinions.

References

- Alderson, J. C. (1995). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp. 71–86). Hemel Hempstead: Phoenix ELT.
- Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, *56*, 137–172.
- Berk, R. A. (1995). Something old, something new, something borrowed, a lot to do! *Applied Measurement in Education*, *8*(1), 99–109.
- Bond, T. (2003). Validity and assessment: a Rasch measurement perspective. *Metologica de las Ciencias del Comportamiento*, *5*(2), 179–194.
- Bond, T., & Fox, C. (2001). *Applying the Rasch model*. Mahwah, NJ: Lawrence Erlbaum.
- Bramley, T., & Black, B. (2008, January). Maintaining performance standards: Aligning raw score scales on different tests via a latent trait created by rank-ordering examinees' work. Paper presented at the 3rd International Measurement Conference. Perth, Australia: University of Western Australia.
- Cizek, G. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement*, *30*, 93–106.
- Cizek, G. (1996). Standard-setting guidelines. *Educational Measurement*, *15*, 13–21.
- Cizek, G. (Ed.). (2001). *Setting performance standards: Concepts, methods and perspectives*. Mahwah, NJ: Lawrence Erlbaum.
- Cizek, G., Bunch, M., & Koons, H. (2004). *Setting performance standards: Contemporary methods*. NCME. <http://www.ncme.org/pubs/items/Setting%20Performance%Standards%20ITEMS%20Module.pdf>. Accessed November 2017.
- Clauser, B., Mee, J., Baldwin, S., Margolis, M., & Dillon, G. (2009). Judges' use of examinee performance data in an Angoff standard setting exercise for a medical licensing examination: An experimental study. *Journal of Educational Measurement*, *46*(4), 390–407.
- Coniam, D., & Falvey, P. (1999). Setting standards for teachers of English in Hong Kong: The teachers' perspective. *Curriculum Forum*, *8*(2), 1–23.
- Coniam, D., & Falvey, P. (2001). Awarding passes in the language proficiency assessment for teachers of English: Different methods, varying outcomes. *Chinese University of Hong Kong Education Journal*, *29*(2), 23–35.
- Drave, N. (2006). The language proficiency assessment for teachers of English (LPATE) as an instrument of educational change. In *Proceedings of the 9th Academic Forum on English Language Testing in Asia* (pp. 18–40). Taipei, Taiwan: CEEC.
- Education Bureau, Government of the Hong Kong SAR. (2010). *Language proficiency assessment for teachers (English language) handbook*. Hong Kong: Government Printer.
- Great Schools Partnership. (2017). *The glossary of Education Reforms*. <http://edglossary.org/standards-referenced/>. Accessed November 2017.
- Hambleton, R., & Jones, R. (1993). *Comparison of classical test theory and item response theory and their applications to test development*. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.690.7561&rep=rep1&type=pdf>. Accessed November 2017.
- Hambleton, R., & Pitoniak, M. (2006). Setting performance standards. In R. Brennan (Ed.), *Educational measurement* (pp. 433–470). Westport, CT: Praeger.

- Henning, D., Hudson, T., & Turner, J. (1985). Item response theory and the assumption of unidimensionality for language tests. *Language Testing*, 2(2), 141–154.
- Hurtz, G., & Auerbach, M. (2003). A meta-analysis of the effects of modifications to the Angoff Method on cutoff scores and judgment consensus. *Educational and Psychological Measurement*, 63(4), 584–601.
- Jackson, T., Draugalis, J., Slack, M., Zachry, W., & D'Agostino, J. (2002). Validation of authentic performance assessment: A process suited for Rasch modeling. *American Journal of Pharmaceutical Education*, 66, 233–243.
- Jaeger, R., & Mills, C. (2001). An integrated judgement procedure for setting standards on complex, larger-scale assessments. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 313–338). Mahwah, NJ: Lawrence Erlbaum.
- Janis, I. (1972). *Victims of groupthink*. Boston, MA: Houghton Mifflin.
- Kane, M. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum.
- Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing*, 21(1), 1–27.
- Luo, G., Seow, A., & Chew, L. C. (2001). *Linking and anchoring techniques in test equating using the Rasch model*. <https://pdfs.semanticscholar.org/6654/23832ea0d72df42e30f1cc7b2975a1bd63e3.pdf>. Accessed November 2017.
- MacCann, R., & Stanley, G. (2006). The use of Rasch modeling to improve standard setting. *Practical Assessment Research and Evaluation*, 11(2), 1–17.
- Morrison, H., McNally, H., Wylie, C., McFaul, P., & Thompson, W. (2009). The passing score in the objective structured clinical examination. *Medical Education*, 30(5), 345–348.
- Qian, D. (2008). English language assessment in Hong Kong: A survey of practices, developments and issues. *Language Testing*, 25(1), 85–110.
- Tannenbaum, R., & Wylie, E. C. (2009). *Using standard-setting methodology for linking assessment scores to proficiency scales: TOEFL iBT and TOEIC assessment exemplars*. Princeton, NJ: ETS.
- Tiratira, N. (2009). Cutoff scores: The basic Angoff method and the Item Response Theory method. *The International Journal of Educational and Psychological Assessment*, 1(1), 39–47.
- Yu, C. H., & Popp, O. S. (2005). Test equating by common items and common subjects: Concepts and applications. *Practical Assessment Research and Evaluation*, 10(4), 1–19.
- Zieky, M. (2001). So much has changed: How the setting of cut scores has evolved since the 1980s. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 19–51). NJ: Mahwah.
- Zieky, M., & Perie, M. (2006). *A primer on setting cut scores on tests of educational achievement*. Princeton, NJ: Educational Testing Service.

Neil Drake is a Senior Manager in the Assessment Development Division, HKEAA. He coordinates the development and marking of the HKDSE English Language and HKDSE Literature in English examinations, as well as the Language Proficiency Assessment for Teachers.

Chapter 15

Misconceptions of the LPATE in the Media: Perspectives on Educational Change



Neil Drave

Abstract This chapter presents research into public perceptions of the value of the LPATE as an instrument of change in English language education in Hong Kong. The chapter reviews media coverage of the LPATE from the years in which the Assessment was most important (2003 to 2007) in the sense of certifying the largest number of serving English teachers. It reviews the press coverage afforded to the LPATE (including letters to the editor, with some opinion pieces), summarising the concerns of the various contributors, but also *critically analysing* the media discourse. The chapter reflects on the nature of good teaching, as well as on the relationship between the worlds of high-impact educational assessment and the popular press. Earlier versions of this chapter were presented at the 9th Academic Forum on English Language Testing in Asia (AFELTA), Taipei, Taiwan (2006) and appeared in the Proceedings (pp. 18–40) as ‘The Language Proficiency Assessment for Teachers of English (LPATE) as an instrument for educational change’; and at the 10th EALTA conference, Istanbul, Turkey (2013), under the title ‘Hong Kong’s Language Proficiency Assessment for Teachers: an impact study’. The PDF is available at <http://www.ealta.eu.org/conference/2013/programme.html>.

Introduction

In its emphasis on standardisation, rigour, security and reliability, as well as the fact that it is a one-off, summative-style test which does not give direct feedback to candidates, the LPATE is a non-developmental ‘high-stakes assessment’ (Diez, 2002, pp. 73–74). It is similar to assessments administered by the USA’s National Council for the Accreditation of Teacher Education and its Interstate New Teacher Assessment and Support Consortium, both large accreditation bodies employing well-defined standards which teachers must meet if they are to achieve certification

N. Drave (✉)

Hong Kong Examinations and Assessment Authority, 12/F, Southorn Centre,
130 Hennessy Road, Wan Chai, Hong Kong
e-mail: ndrave@hkeaa.edu.hk

© Springer Nature Singapore Pte Ltd. 2018
D. Coniam and P. Falvey (eds.), *High-Stakes Testing*,
https://doi.org/10.1007/978-981-10-6358-9_15

323

(Diez, 2002). LPATE resembles the ‘competency tests’ (of literacy and numeracy) which are administered to all prospective teachers in the USA and elsewhere, normally on entry to and exit from teacher education and/or certification programmes (Brookhart & Loadman, 1995). In such accreditation initiatives, assessment may be regarded as one tool for effecting improvements in teaching and (consequently) of learning, particularly within what Moore (2004) calls the discourse of the teacher as ‘competent craftsman’.

Other chapters in this volume, particularly in Section IV, describe data-based research on the direct educational impact of the LPATE, from the perspectives of education policy and language pedagogy. This chapter describes research which sought to address the question of the impact of the LPATE indirectly, through the lens of media discourse: how has the assessment been presented by and in the media, and what have been the possible consequences of this on whether the LPATE has been able to promote educational change?

In the following section, the data and methodology of the study are described. The discourse framework of Moore (2004) is used to place the data in the context of wider debates about teachers and teaching. The chapter ends with brief comments on the relationship between assessment and educational impact, commenting on the connections and contradictions between the two areas.

Methodology

In seeking to analyse the impact of the LPATE, English language newspaper items published between December 2003 and July 2012 were scrutinised, a period covering twelve LPATE assessments (see Appendix ‘[English Language Newspaper Articles December 2003–July 2012](#)’). This period was chosen because it was felt that only after several iterations would the ‘true’ impact of the assessment have emerged, hence the decision to neglect pieces published within the first two years of the LPATE’s operation. The majority of items come from 2004, 2005 and 2006, which was when the candidature was at its highest. As described in Section I, the LPATE was originally scheduled to run for five years, ensuring that all teachers who needed to do so would have time to attain the required ‘benchmark’ level in all papers before the beginning of the 2006–07 school year (in September 2006) [Note 1]. As a result of policy decisions, the last test in the original format took place in February 2007, after which the assessment was revised and offered annually to new teachers only. Since 2012, there has been no significant media discourse about the assessment, which suggests that it is no longer controversial and has been accepted by teachers, a conclusion corroborated by the data presented in Section IV of this volume.

English language media reports are widely recognised as the most authoritative source of information for many in the Hong Kong educational community, and many people use them if they wish to reach an international audience or set an agenda, so these constitute a source of useful and reliable information on public perceptions (see Habermas, 1966, 1991, 1993). Media is a lens not a mirror. Naturally, the inclusion

of Chinese language data would have made for a more complete study but this was beyond the scope of this project.

The data were gathered using WiseNews, a widely used news search service, and contained letters to the editor ($N = 23$), opinion pieces ($N = 9$) and hard news ($N = 22$) from five prominent Hong Kong English language newspapers and one online news source, as well as two questions posed to the Legislative Council and the response to them. Of the 55 pieces, seven were written by (self-identified) teachers, nine by academics, 23 by journalists, two by civil servants working for the Hong Kong Education Bureau and two by the Manager (formerly the 'Subject Officer') at the HKEAA responsible for running the LPATE. Fifteen were classified as generally positive in their attitude towards the assessment, while 27 were negative and 13 were balanced, non-committal or not evaluative in nature. An academic article on the impact of the LPATE (Glenwright, 2002) supplemented these data.

The analytical approaches adopted were broadly within the traditions of content analysis, in which categories are created and occurrences of items counted, and critical discourse analysis (Fairclough, 1995), in which presentation style is analysed for linguistic and pragmatic features which might influence the way the discourse is interpreted.

Change Models

Many models of how behaviour can be changed to emphasise the importance of affective factors, such as belief, in the change process (e.g. Fullan & Stiegelbauer, 1991). It is possible to get a sense of such factors by looking at editorials, letters and news items, the documents of the 'public sphere' (Habermas, 1966, 1991, 1993). According to Habermas, in eighteenth-century Europe, the market economy gave rise to a formalised notion of public opinion which provided an oppositional force to the hegemony of the (parliamentary) monarchies. Groups of citizens gathered together, often in coffee shops, to talk about matters of public importance and in this way formed a sphere of influence which allowed the issues of the day to be freely debated. There were four distinctive features of this (idealised) sphere: (1) individuals came together, (2) around issues of general interest, (3) without concern for social status and (4) in order to achieve rational consensus by means of critical discussion (Calhoun, 1992; Fraser, 1993). The role of this public sphere was to 'amplify the pressure of problems, that is, not only detect and identify problems but also convincingly and influentially thematise them, furnish them with possible solutions, and dramatisise them in such a way that they [were] taken up and dealt with by parliamentary complexes' (i.e. the real decision-making bodies; Habermas, 1966, p. 359).

There has never been a perfectly unbiased and egalitarian sphere, as Habermas later admitted, and the coffee shop has been supplanted by formal and informal media organisations and online communities. Newspapers (in print and online) now provide one of the most important forums for public debate, but they also have to

sell themselves, and they do this by various techniques, including sensationalising stories. Nevertheless, the prominence and reception of ideas in the sphere act as a general barometer of public opinion, which in turn forms an important check and balance on the power of governments. The HKSAR Government is a case in point. It is aware of criticisms of the lack of a popular mandate for its policy makers and that the Legislative Council (which is only partly elected) is unrepresentative of the views of the general public. It has therefore sought the opinions of ‘the people’ directly on a number of controversial issues in recent years, including such matters as harbour redevelopment and smoking in restaurants. Indeed, that it is not elected probably makes the government more sensitive to public opinion, even to the extent of wishing to directly influence public opinion polling [Note 2]. Lacking recourse to the flexibility and freedom offered by a popular mandate, it must participate in a day-by-day process of consensus building. It is therefore probable that the Education Bureau (EDB), which administers education policies, is fully cognisant of, and sensitive to, the public reception given to its policies and regards public comment as a fair reflection of the success—or otherwise—of policy implementation.

There are alternative theories of how the media influence ideas in society. Herman and Chomsky (1988), in their classic study, promote and exemplify a ‘propaganda’ model in which the media ‘inculcate and defend the economic, social and political agenda of privileged groups that dominate the domestic society and the state.’ They do this by deciding what gets into the media in the first place, what makes news, as well as by the way stories are framed and the amount and quality of the information which is allowed to be printed. The argument is that this is done consistently, not to reflect the flow of issues but to deliberately exclude certain perspectives and prevent some voices from being heard.

This may seem to be an extreme opinion but, at the very least, it is clear that the media select what is to be published (i.e. it exercises a ‘gatekeeping’ function; see Watson, 1998 for a summary); *Letters to the Editor* are an example of this practice. In covering certain aspects of a topic at the expense of others, it is possible to influence which aspects of the issue become prominent and which are ignored by the public (Watson, 1998, pp. 112–113).

It has been proposed that decisions about what kinds of stories make news depend on the notion of news *values* (Watson, 1998). A story is more likely to be published if it has certain characteristics (Watson, 1998), for example:

- **Frequency**—stories which have an appropriate timescale for news production are favoured.
- **Amplitude**—the more dramatic or impactful a story, the better.
- **Unambiguity**—issues which are simple, or which can be simplified, are favoured.
- **Negativity**—bad news is more newsworthy than good news.
- **Personification**—human interest stories are newsworthy.
- **Eliteness**—the opinions of society’s elite are more newsworthy than those of other people.

Once a news story has been chosen, issues of distortion and replication come into play (Watson, 1998, p. 122).

The relationship between the media, public opinion and reality is complex and contested. The media provide an outlet for the public to express their views, but it is up to the media themselves to decide which opinions get expressed, how these are edited and what 'news' stories appear. As shown in the Findings section below, these factors have influenced what has appeared in print about the LPATE, with the esoteric nature of assessment being an additional, complicating factor.

Findings

The findings are presented below under thematic headings which relate to (1) whether the LPATE is perceived as an isolated initiative or part of a coordinated package of educational reforms; (2) whether the assessment is felt to have had positive or negative consequences; (3) the quality (accuracy, level of sophistication) of the contributions to the media debate; and (4), from a critical perspective, the linguistic techniques used to present information. The remarks do not consistently distinguish between hard news and other kinds of writing except obliquely in referring to writers of opinion pieces and *Letters to the Editor* as 'contributors'. It is noteworthy that the former Permanent Secretary for Education and Manpower, Fanny Law, is on record as saying that the LPATE is an example of a reform which was well planned ([14]). [Note 3]

Attempting to Understand the Intended Impact of the LPATE

What was the intended impact of the LPATE? As described in Section I, the Assessment was introduced on the recommendations of the Education Commission which, in its Report Number 6 (December 1995), stated the aim of 'ensuring that new teachers are competent to teach through the chosen medium of instruction' (Education Commission, 1995, Section C2). The Advisory Committee on Teaching Quality (ACTEQ), which was charged with advising on the implementation of the policy, later narrowed the scope of the initiative to include only language teachers. The LPATE was one of a number of measures intended to increase professionalism and bring about an improvement in classroom practice. Perhaps the most important of these was the requirement that all new teachers should be subject trained university graduates. The data in the present study suggest that there may have been confusion about the large number of educational reforms and about the relationship between the different initiatives. In many articles, the LPATE is mentioned in tandem with other education reforms: some of these were indeed directly related in reality (e.g. school-based assessment, [13], quality assurance [15]), while others were not (e.g. the launch of a new degree in a teacher training institution [46], class size, subject-matter

training [12], [44], school structure, medium of instruction [51] and school closures [13]); overall, correspondents display a confusion over the relationship between the LPATE assessment and other education policies. In some pieces, the LPATE is presented as having to bear the brunt of English language reform on its own, which was never the government's intention, and some writers therefore display (understandable) disappointment that the LPATE 'has not achieved anything educationally substantive' ([2]).

Some of the blame for such misunderstandings certainly rests with policy makers. There could have been a more clearly articulated rationale in the initial design and implementation stages of the LPATE, for example. The preparatory documents for the assessment do not set out precisely how reform or improvement in English was to be accomplished, or in what framework it was to function. The assumption that there were shortcomings in language teacher performance which could be addressed by benchmarking were also not well articulated in these original documents and the relationship between the LPATE assessment and other ways of achieving the benchmark might have been confusing to many.

The English Language Benchmark Subject Committee Report (Coniam & Falvey, 1999), the most important pre-implementation document, refers in one paragraph to positive backwash, citing the Australian ELSA experience (Australia Language and Literacy Council (1996: 20–21), but does not otherwise address the mechanism for teacher improvement. One can assume, however, that underlying the rationale for the LPATE was that it would have an impact in one of the following areas: if English teachers displayed good (enough) English, this would promote/constitute good English teaching; the assessment might weed out the teachers with poor English and so prevent bad modelling/teaching practice. These assumptions may logically lead to different solutions, some developmental and some purely criterial, and did in practice lead to a policy of allowing teachers to demonstrate proficiency in a number of different ways, including the LPATE. This proliferation of means and ends could have led to confusion in the public mind.

One reason for the lack of detail on the above assumptions is that the relationship between the language of the teacher and that of the student, as well as the teacher's awareness and understanding of the language which they are teaching, were at the time of LPATE's introduction unproven. The evidence that teacher language awareness (TLA) is an important variable in the teaching equation was only then beginning to be published (e.g. Andrews, 2001). Likewise, there were more far-reaching educational discussions (still ongoing) which sought to define what good teaching is, plus a number of economic considerations to contend with, consequences of the Asian financial crisis of the late 1990s, and therefore a number of different paradigms were (and still are) extant. It is therefore understandable that confusion should arise about the functions of the LPATE within the broader educational change process.

Positive and Negative Impacts: Different Points of View

About half of the pieces are critical of the LPATE. Negative comments include accusations that the assessment takes teachers away from their core job of teaching, adds to their workload or that it is a bureaucratic and expensive assessment (e.g. [4]). One recurring theme in the data is that educational change is burdensome, so the more changes there are, the greater the burden on teachers. LPATE is often discussed in the context of this ‘burden’ (e.g. [20]), although there is little discussion of precisely how this one-off assessment is actually burdensome. There are many references to the assessment as contributing to an increase in teacher workload ([8], [12]) although once again details are scant. This opinion was expressed in print as recently as 2012: ‘An example [of an educational initiative that is out of touch with frontline educators] is the complex LPATE, which has unreasonably increased the burden and workload for language teachers in Hong Kong’ [i].

There are several areas in which correspondents are divided: whether to trust teachers, whether certification is needed, whether to trust the results and whether the scope of the LPATE should be expanded.

There was a feeling of ambivalence towards teachers, even among teachers: it is felt that some teachers are not very good, but they are also victims of government policy ([2]). Some correspondents feel that LPATE has ‘cut the dead wood’ ([4]) but that there are still some ‘low or middling range teachers in the target language’ ([4]) who may still harm students. Several writers feel that the test should be extended to those teachers who are teaching other subjects through English or that it should also test teaching skills.

There is also ambivalence towards certification, and even though this is a high-stakes assessment, there was an unwillingness to acknowledge that important decisions should be based on its results. Assessment results are felt to be untrustworthy since test performance does not reflect the real abilities of teachers ([18]) and even those who pass cannot be trusted ([41]). There was a reluctance to label teachers as failures: ‘Even though they have not met the requirements, it does not mean that their language skills are poor’ ([26]). Some common stereotypes about language teaching are transferred to the world of testing, for example that employing native speakers of English as oral assessors guarantees the quality of the assessment ([1], [4]). In fact, both native and non-native speakers are employed in the LPATE assessment process and the assessment statistics show that it is very reliable, with good inter-rater reliability (see Urmston, this section). There is thus a general mistrust of the assessment on what might be called ‘technical’ grounds (Filer, 2000), that is, on issues such as reliability and generalisability. Some writers ascribe more (negative) washback influence to the assessment that is likely in real life. For example, two writers assert that the error correction and explanation part of Paper 2 (Writing) will encourage more error correction in the classroom and prevent innovative pedagogy ([31], Glenwright, 2002).

Undoubtedly, teacher correspondents are nervous about change, and, in particular, they ‘are apprehensive about the spread of benchmark culture to include an assess-

ment of [their] own language ability' and fear a loss of face if they fail (Dowson, Bodycott, Walker, & Coniam, 2000, p. 19).

The mistrust of certification is understandable since this is a relatively recent worldwide trend. It has been accompanied by the imposition of standards of accountability and the encroachment of commercial values which have traditionally been the province of the business world on other spheres which were traditionally above such concerns. Teacher accountability is a worldwide trend. Teachers in Florida are now paid according to performance, for example and America's *No Child left Behind* act, with its public dissemination of classroom test results and school rankings, made it increasingly common for teachers to be judged on learning outcomes. The UK's 1988 Education Reform Act, which led to the National Curriculum and the system of National Vocational Qualifications in the UK, was an early (and problematic) leader in this trend (see Drave, 1996 for linguistic problems with the NVQ documentation), which has also reached New Zealand (Priestly & Higham, 2002) and elsewhere.

It may be true to say that educational change of any kind and under any context will spark such concerns, (Guthrie, 2011; Köksal, 1995; Zimmerman, 2006). For instance, similar reactions to the ones found here greeted the introduction of the UK's National Curriculum (Broadfoot & Pollard, 2000, p. 15). However, the economic climate in Hong Kong in the early 2000s, in which cost-cutting measures saw salaries fall and budgets tighten in all sectors of society, was a particularly difficult one in which to introduce stringent certification measures. Additionally, an ageing population, falling birth rate and increasing number of education providers, particularly at sub-degree level, combined to make Hong Kong education more of a buyers' market than ever. Many of the concerns expressed in the data related to these wider concerns; of teachers fearing for their jobs and many commentators wishing to uphold (or improve) English standards in the face of economic and demographic challenges.

The Quality of the Debate

Perhaps it is wrong to expect too much from media discourse, which has to conform to news values and fit into the constraints of limited column inches, and unfair for an assessment insider to comment on the contributions of those who may have only a passing acquaintance with the LPATE. However, since the articles were intended to be in the public sphere, they are surely a legitimate topic for comment and criticism. In general, however, what emerges is that the debate seems to be marred by misunderstandings about assessment, poor argument and some failures of reporting and editing.

When making general claims about teaching or language standards, there is a reliance on purely anecdotal evidence or on none at all: 'Walk past almost any school building in HK...' and you will learn what happens in all classrooms, apparently ([4]). Some correspondents unashamedly acknowledge where their information has come from: 'anecdotal evidence suggests that not everyone employed to teach English can communicate effectively in the language...' ([19]). Again, the common

misperception that the ‘standard of English’ is falling is repeated many times, for example in [19], which claims that ‘most analysts maintain’ that this is true. It would be interesting to know who these analysts are and how the writer was able to gather their opinions.

The aims of the test are not well understood, with many under the misapprehension that it is about teaching (e.g. [7], [19], [49]). Newspaper headlines are particularly misleading in this respect: ‘Poor English teachers must mind their language’ ([22]). The LPATE is an English test, hence largely cognitive in nature, with only one of its five components undertaken in the classroom setting. It is therefore (at best) only a ‘proxy measure of teaching effectiveness’ (Rich, Barcikowski, & Boyd, 1995). The Director of the British Council points out this important distinction. She asserts that the LPATE is a ‘useful diagnostic benchmark but it won’t in itself improve English standards in Hong Kong’ ([5]).

Perhaps more understandably, given its relative recency as a testing mode, there are misunderstandings about the nature of standards referenced assessment (a term often used synonymously with criterion-referenced assessment), such as the fact that standards are held constant from test to test (through pre-testing of an anchor paper, *inter alia*; see Drave Chap. 15). A common misconception is that one test is more difficult than another and that the difficulty of the test is significant for the final candidate outcome ([35], [49], [50]): ‘Top candidates teaching at the exclusive DGJ School said that the test was too hard’ ([9]). The standard is also felt to be too high: ‘Good teaching at primary and lower secondary levels doesn’t need near-native proficiency’ [49]; ‘While it is essential for language teachers to have a sound foundation of (*sic*) their Chinese and English language skills, it is unnecessary to make the LPAT excessively difficult for language teachers, in order to establish the reputation of the LPAT’ [i]. Also, some commentators, without referencing the stability of standards, wrongly think that performance should be stable from round to round, even though, as an open-entry proficiency assessment, the test population may be different each time ([18]).

One un-named academic claims that the test is ‘flawed and needs overhauling’ because the pass rates for listening are different from one assessment to another and because of differences in the pass rates for Reading and Writing ([34]). In fact, research has found that there is no consistent, reliable, significant correlation between productive and receptive language skills for Asian learners (Hirai, 2002; Poedjosoedarmo & Hsui, 1996). Some writers simply do not know anything about the test or how it is put together, as evidenced by the request to make sure papers are ‘standardised’ (whatever this means) ([42]). A norm-referenced mentality is also in evidence: if so many candidates fail, this must be a bad test ([49]).

Some writers claim knowledge which they cannot possibly possess, e.g. the assertion that half the failures are ‘veteran teachers who know how to manage class discipline’ ([8]), or that now, some years after inception, only the weaker teachers are left to take the test and this explains the variability in performance ([37], [25]). One correspondent refers to the concept of a ‘bare pass’ ([18]). However, this does not really apply in the (standards referenced) LPATE since a candidate’s performance is simply assigned to one of five levels. It is possible to be a borderline candidate in

two of the five papers, where a cut score is used to determine proficiency, but this raw score has no real utility outside the standard-setting process and neither it nor the cut score are released to candidates or the general public.

Some pieces focus on only part of the assessment as if it were the whole. The Writing Test (Part 2) is typically singled out for attention as it requires candidates to demonstrate their language awareness, which is felt to be very demanding, and it is this paper which often has the lowest attainment rate. The assessment apparently ‘do[es] not take into account the fact that a band-three teacher [Note 4]... may require other qualities apart from an ability to explain grammatical constructions’ (19), says one writer, ignoring the remaining tasks in the other papers, none of which test the stated skills.

Some writers expressed the opinion that the Writing Test was irrelevant to the job of language teaching, but evidence to the contrary is now emerging. Andrews has conducted extensive research into the impact of teacher language awareness (TLA) on the classroom practice of Hong Kong teachers (e.g. Andrews, 2001, 2007). He has come to regard it as a kind of filter which good teachers are able to use flexibly in different circumstances. Andrews’ three conclusions are that (1) when teaching grammar, teachers who know it better, and can explain and reflect on it better, may teach it better (i.e. TLA is necessary but not sufficient for good teaching); (2) better TLA may result in better classroom input; and (3) better TLA results in greater willingness to engage with language-related issues, which is itself a positive thing for the classroom context (Andrews, 2001, p. 86). The LPATE may be able to screen out Andrews’ ‘Teacher C’ and ‘Teacher D’ types, those without the necessary knowledge to engage with students on content-related matters, even if (as in the case of C) the teacher is willing to do so.

These distinctions are obvious to correspondents who do have some knowledge of assessment (usually tertiary teachers) and there are many complaints about the standard of hard news reportage, for example the negative and misleading coverage (e.g. [37], [39]). These contributions form a counterpoint to those from teachers and the general public which, in their lack of knowledge about testing and the mechanics of the LPATE, rely on personal and anecdotal evidence to support their assertions.

Drave (2006) discusses a range of different but related issues which arose from the data, such as where the promotion of commercial and other interests, were discussed. The reader is referred to that paper for more information.

A Critical Approach to Hard News Coverage

In the 22 hard news stories, some of the common features of newspaper discourse are in evidence. There is also evidence of the operation of news values, for example selectivity (choosing to report only bad news), and slanting (using emotive negative vocabulary), which may have influenced public perceptions of the LPATE. Some of the key features (see, for example <https://www.thebalance.com/hard-news-how-does-it-differ-from-other-types-2316022>; Reinemann, Stanyer, Scherr, & Legnante, 2011) are:

- The use of *punning* headlines, which often sacrifice accuracy for rhetorical effect, playing on the novelty of teachers being tested (rather than doing the testing): ‘language tests for teachers a failure’ ([2]); ‘2000 teachers fail to make the grade’ ([11]); ‘Poor English teachers must mind their language’ ([22]); ‘Language teachers fall below mark’ ([26]); ‘Teachers on the bench as they fall short of the mark’ ([28]); ‘Teachers in English struggle’ ([42]); ‘Teachers must mind language’ ([51]).
- *Presupposition* that there should be higher pass rates, indicated by words like ‘just’: ‘Just 29% of teachers get pass in English test’ ([25]); ‘Just 30% of teachers... passed the writing paper, a marked decline ... ([18]); ‘Only 35% of 943 teachers passed the writing test’ ([21]); ‘Just 29% of teachers get pass in English test’ ([25]). Declining English standards are also taken for granted in comments which address the ‘underlying reasons for the declining standards of English in HK’ ([24]). Journalists are not afraid to comment on what teachers should be able to do: ‘Some candidates even failed to write a grammatically correct and complete English sentence’ ([42]). Factual distortions abound, for example the claim that ‘only a handful of teachers who appealed for a reassessment of their *low* marks... were successful’ ([47] emphasis added); in fact, any candidate can appeal whether they feel that their ‘marks’ are ‘low’ or not. There is therefore an evaluative, judgmental element to the reporting of the ‘hard’ news which suggests disappointment in the candidates’ performance and in the attainment rates, but nowhere is this explicitly referred to or made a topic for discussion so that it might be challenged.
- *Negativity* in choosing to highlight failure and negative emotions: ‘Third of English teachers fail key test’ ([40]); ‘Teachers fail to get higher marks’ ([47]); ‘Dubious English tests’ ([49]); ‘A to Z of reforms that have educators hopping’ ([15]); ‘Language teachers fall below mark’ ([26]); ‘Job fears for 1500 teachers’ ([6]). Where articles report on results selectively, they almost always give only the results for the papers with the *lowest* attainment rates (usually Speaking and Writing).
- *Dramatisation* of results, consequences or emotions for sensational purposes using strong lexis, as in the following (emphasis added): ‘English teaching skills *plummet*’ ([41]); ‘Educators *demand* tests’ disclosure’ ([50]); ‘results... dropped *sharply*’ ([23]). It also seems that newspaper editors are trying to create through rhetorical means a simple conflict between parents and teachers by quoting ‘opposing sides’ ([9]) in the ‘...language test *crisis*’, which is looming as ‘teachers fail crucial test’ ([18]). They even claim that ‘Language tuition in schools [is] ‘under threat from test’ ([23]), which is a surprising twist.
- *Non-specific sources; restricted range of sources*: only a narrow range of sources is cited. Sometimes the source is ‘According to educators’ ([21], [44]); almost all other comment is ascribed to one of a small group of sources, including Fanny Law, representatives from the JET Circle and Education Convergence (education pressure groups), the Professional Teachers Union and the Alliance of Parent Teacher Associations. Close to a quarter of the articles did display a balance of points of view, however, citing both opponents and supporters of the LPATE.

Based on the evidence cited earlier, it seems that there are some problems with the public sphere in so far as it is able to cope with debate in this arena, which may

mean that the community recognition and understanding of the LPATE as a (positive if limited) change instrument may be diminished. As stated earlier, the debate is marred by misinformation, lack of information, poor arguments and ignorance of assessment principles and practice. Information is given and contributions are made by a small number of people. The reporting is heavily slanted to the negative and sensational. In short, the LPATE is treated as just another news story.

Assessment issues become conflated with both the causes of teacher stress and the Government's education policy initiatives without the relationships between them really being teased out. The fact that different policy initiatives were introduced at the same time, initiatives which are usually felt to presuppose contrasting assessment goals (as we shall see), seems to be a source of confusion about how they will impact teacher work, how they will relate to each other and how each is supposed to achieve its ends. Those who choose to write letters often do so primarily to promote their own agenda rather than to contribute to meaningful debate (Drave, 2006).

Common stereotypes and misunderstandings about English standards, teaching and teachers, testing and the LPATE and its aims occur again and again. Filer (2000, 2) identifies two distinct categories of discourse about assessment, the 'technical' and the 'sociological'. In the 'technical' discourse category, the ends of assessment are taken for granted and the discussion concerns the means. Many of the letters and articles in this study discuss these issues, commenting on matters of reliability and standards. However, given the paucity of information and the constraints of the format, these discussions remain at a superficial level. The matter of misinformation is troubling but not surprising, given the esoteric nature of assessment and the speed with which newspapers are produced, which leads to shortcuts in research. In such circumstances, the influence of public discourse is bound to be limited to rehearsing the views of particular insiders or the personal experiences—the 'lifeworld' in Habermas' terms—of the 'victims' (who have no way of transcending the personal). However, the sharing of personal experiences of the subjective experience of the testing process (part of the lifeworld), for example may indeed be influenced on assessment policies and practices (part of the 'system'). It may be too easy to dismiss such contributions to the debate as being irrational and therefore irrelevant or powerless to effect change; the important question is 'whose personal experience is valued'?

Filer's second discourse category is the 'sociological' discourse category in which the relationship of assessment to matters such as social reproduction and control are at issue and in which the social context of assessment is of paramount importance. The contributors to a volume edited by Filer (2000), for example regard assessment as a form of social control, 'the archetypal representation of the desire to discipline an irrational social world in order that rationality and efficiency could (sic) prevail' (Broadfoot & Pollard, 2000). The data in this study occasionally touch on such issues but there is little real engagement with them, which is understandable given the nature and sources of the data and the expertise and interests of the contributors.

The Broader Issue of Impact: Teaching Quality

There is a widespread consensus that the quality of teachers is a key ingredient in the learning process and that what teachers do and say in the classroom is very important, so it would be logical to assume that any initiative which seeks to improve teacher language proficiency would promote better teaching. However, there is widespread confusion over what makes a good teacher and this leads to confusion about the relationship between teaching, teachers, assessment and education policy. Moore (2004) has identified three discourses about teaching, in which teachers are ‘competent craftspeople’, ‘reflective practitioners’ or ‘charismatic subjects’. The last is often portrayed in the media as eccentrics or as being in conflict with the teaching establishment (Robin Williams’ character in *Dead Poets’ Society* is an example of a charismatic subject).

In the first, positivist paradigm, teachers (as ‘competent craftspeople’) must be able to demonstrate certain skills or knowledge. For example, from the world of English language teaching, the well-respected and very powerful TESOL organisation has a set of standards which teacher training programmes need to adhere to when they assess trainee teachers (TESOL, 2010). There are five domains (Language; Culture; Planning, implementing and managing instruction; Assessment; and Professionalism), each of which has a set of performance indicators. Such a standard-driven approach has as its aim to codify and standardise the content of teacher training programmes so as to set clear goals for such programmes and allow teacher educators to distinguish between those who are suited to the teaching task and those who are not. Understandably, its definiteness and comprehensiveness, plus the fact that the skills are assessable, make this approach popular with governments and other policy makers (Moore, 2004, p. 14). Assessment of teacher competence in this paradigm is a necessary part of the overall framework of educational policy and, in so far as it weeds out those without the necessary skills, can serve the needs of educational improvement (in this very specific sense).

The need to demonstrate certain linguistic skills in the LPATE assessment, the classification of different levels of performance according to performance descriptions, and the fact that failure in the assessment disqualifies people from teaching in Hong Kong, are all features which presuppose this ‘skills’ view of teaching. The data in this study show that some correspondents subscribe to this view and welcome the ‘getting rid of dead wood’ approach to raising standards. Others, however, disagree and shy away from the logical conclusion to such an assessment, that those who fail it are not fit to teach. One objection is that even existing teachers have to pass the test, which calls into question the quality of Hong Kong’s previous teacher preparation programmes and the system of which they have been a part, a criticism which many are unwilling to make.

In contrast to this objectivist approach, the ‘teacher as a decision-maker’ view regards the teacher as someone who engages in more complex and less easily describable classroom activities and is able to reflect upon such activities in order to improve learning (Moore, 2004, 201; see Glenwright, 2002 for a critique of the LPATE from

this perspective). This is probably the most influential view among teacher educators today.

Clear precedents and underpinnings for the reflective view of teaching/teachers can be found in the field of applied linguistics and, more specifically, of English for Specific Purposes. The key idea is that professionals distinguish themselves as a group by the ways in which they think and talk, by their discourse and their ways of interacting. Logically, then, professional ‘training’ of skills, which traditionally focuses on the observable, physical skills of manipulating objects or directing people, should give way to the (by definition) higher level ‘education’ of teachers. Becoming a professional teacher or other professional is thus more than just demonstrating a way of presenting subject content; it includes ways of thinking, problem-solving and fundamental attitudes towards the job and the people in it. The education of an individual is not something that can be accomplished mechanically and quickly but is an extensive process of understanding and reflecting on professional practice and the beliefs which underlie them (Richards, Gallo, & Renandya, 2001). These ideas were popularised in Hong Kong by, among others, Jack Richards, who was professor of English at City University during the 1990s and introduced these ideas to his students, many of whom went on to become English teachers.

From a broader sociological point of view, this seemingly contradictory yet co-existent duality—simultaneous emphasis on standardisation and on reflection—is perhaps explainable by the conditions of ‘high modernity’ in which we live (Giddens, 1990). In modern society, information which was once the preserve of the specialist (doctor, lawyer) is increasingly easy to obtain by all, so there are fewer reasons to seek professional advice for its content alone. One seeks such advice because it is mandatory (e.g. for medical insurance purposes), uncodified (as is much of English law) or because there are additional benefits to doing so (e.g. the persuasive skills of a trained barrister are needed in court). Joining communities of professional practice is arguably therefore less about possessing a certain body of knowledge than about possessing appropriate qualifications or being able to talk to other practitioners in a particular way (i.e. sharing a discourse).

Research on change in the ELT context, from the reflective theoretical perspective and as perceived by teachers, has demonstrated that it is a multifaceted process (Richards et al., 2001) and relies (at least in the constructivist paradigm) on the development of personal theories of teaching. In their research on English teachers in south-east Asia, Richards and his colleagues discovered that it was training courses which gave the greatest stimulus for change, followed by student feedback (2001, p. 9). They cite research which identified several catalysts for change, including:

- Dissatisfaction with the current situation,
- A change in the teaching context,
- Life changes or personal growth, which led to professional development
- Conflict between the teacher’s new beliefs and their practices.

Only the second of these is an external factor, the others are internal, and even the second will not succeed unless the teacher does something to respond to the

changing context. Interestingly, external assessments did not figure at all in the teachers' responses. The authors of the Broadfoot and Pollard's (2000) ELT study admit to methodological flaws which might have influenced participants' responses, but if we accept their findings, and the paradigm, the appropriate question to ask is: could assessment initiatives succeed in bringing about change by impacting teachers' beliefs about what they are teaching and the way they are teaching?

One useful psychological theory is implied in the last of the above points, relating to conflict: 'cognitive dissonance', a theory developed by Festinger in the 1950s (Festinger, 1957). This theory suggests that beliefs can be changed if there is a conflict (dissonance) between existing practice and new information. For example, if a teacher were to learn that her pronunciation of a particular sound was wrong, then she might be prompted to pay more attention to this and might be able to change it. In this way, assessment could be an instrument for change, although one might question how long-lasting such a change might be and how it would actually be effected in practice. The outcome is more likely to be a greater sense of awareness that professionalism is required, professionalism which would manifest itself in certain behaviours (this is noted in the analysis of data in the research exercise described in Section IV). There might also be improvement if a teacher were to take a test several times or were to actually revise or practice in preparation for it.

Some correspondents in the current research promote the idea of the reflective practitioner, and object to the LPATE on the grounds that it does not contribute to teacher development, either because it is by nature unsuited to this task (which is true, since it was never designed to be developmental in this sense) or because it takes up time that would otherwise be used for such development. However, we should remind ourselves that teaching (and teacher training) in Hong Kong takes place in the teacher's second or third language and that many schools function exclusively in Chinese. This language mix, in which English is often confined to the performance of certain designated and limited functions in the classroom, is a fly-in-the-ointment for reflective practice in the language; the intellectual quality of the reflective approach is likely to be tempered by a need for verbal routines which achieve class management needs and give teachers the guidance they need to get through the teaching day. It is also likely that only those teachers whose grasp of English is very secure will be able to meaningfully reflect on their practice in that language.

Ironically, those parts of the LPATE which were most criticised were those which asked that appropriately 'teacherly' ways of talking be displayed, for example the task of 'explaining language matters to peers' in Paper 4. This is very similar to Moore's 'intra-professional verbalised reflections' (2004, p. 105), a sharing of classroom experience with peers which he identifies as one of five sites of reflective practice. The key difference, then, is not in *what* the LPATE asks of teachers, but *how* and *over what period of time* it does so. If the assessment were to be more extensive and give feedback, it could contribute to reflective practice. And, of course, there is no suggestion that benchmarked teachers do not have further to develop as, by definition, the LPATE is only concerned with English and only establishes minimum proficiency. Therefore, it is possible, even desirable, for initial assessment and reflection to co-exist within the education system. However, current education policy, which attempts

to emphasise *both* minimum entry standards and developmental support for teachers, can only be reconciled if stakeholders are prepared to resist the temptation to choose between these approaches and if the education system is able to support them both.

The third of Moore's paradigms, that of the charismatic individual, is unlikely to be amenable to change through assessment as, by definition, inherent, personal factors are paramount here. This is the popular, mythical conception of the teacher as a source of inspiration, whose endless love and concern for bringing out the best in students often opposes them to the formal mechanisms of the school and educational system. The importance of personal factors to the trainee teacher is supported by research suggesting that trainees often learn very little from their training since their own school experience, plus what they bring from the 'real world', act as a filter on what they learn in class, although some teachers are willing to acknowledge this and challenge their preconceptions (Moore, 2004, pp. 15–16).

What these three discourses (craft, reflection and charisma) have in common is an emphasis on what teachers *do*. In the first, the assumption is that there is a cut-off point at which one is able to teach, and so assessment of some kind will be very important because it determines and implements this cut-off. Educational change is more likely to be about changing the formal structures which support and implement policy, and seeking consensus about the required standards; in the second, assessment is less likely to be 'high-stakes' and more developmental since good teaching is something which one works towards and continues with over time, not something which needs to be demonstrated *a priori*. Educational change is more likely to be teacher-focused. Acceptance of change will not, *ipso facto*, change classroom practice however, since this is a complex mixture of belief and skills which is influenced by many factors in addition to policy and only happens over a considerable period of time. The third paradigm is likely to scorn assessment (of both teacher and students) as being at best irrelevant and at worst destructive and will not see the need for systematic educational change. Each of these views was expressed in the data in this study; some writers bemoan standards and demand more tests of skills; others hint that teaching is more than skills, or cannot really be trained at all as it relies on personal qualities such as motivation and effort; still others ask for more time for teachers to reflect on their own practice.

Conclusion

There are a number of limitations to the present study, particularly in the matter of the range of data, which is here restricted (more or less) to the print media in English. Likewise, debates about good teaching and improvement always throw up different definitions (discourses), some of which are mutually incompatible, and these need to be further explored. If we cannot decide what good teaching is, how can we promote it?

Assessment is a poorly understood and esoteric part of the education system and is often seen as antithetical to the predominant discourse of reflective practice

(at least in universities). The co-existence of different discourses of teaching in the educational community has given rise to a lack of clarity over the aims of the LPATE, and consequently to expressions of disappointment that it has not been able to change classroom teaching practice. Hong Kong is still coming to terms with the need to reconcile the various key issues: the community's desire for high standards, the demands of increased professionalism and certification, economic and social realities exacerbated by the foreign language environment, as well as the worldwide movement towards teachers as reflective practitioners who are also accountable (sometimes financially) for learning. The LPATE is, above all, a site of contestation in which the broader educational issues are problematised and played out; it is this which has made it important for recent education policy in Hong Kong.

Notes

1. For further details, see Urmston (2002, 2003).
2. There were allegations in 2000 that former Hong Kong Chief Executive Tung Chee Hwa tried to interfere in the work of a pollster from Hong Kong University.
3. Numbers in square brackets refer to individual articles or letters. See Appendix A 'English Language Proficiency Standards 2015' in Chap. 16 for details.
4. Students are allocated to different Secondary schools on the basis of their Primary school assessment results. There are three 'bands' of Secondary schools, with band three being the lowest, i.e. offering places to the students who have performed least well.

Appendix: English Language Newspaper Articles December 2003–July 2012

No.	Article title (date of publication)	Type	Source	Stance	Writer
<i>Post-LPATE revision articles</i>					
(i)	Reflecting on Hong Kong's education sector since handover (4/07/12)	OpEd	China Daily (HK Edition)	–	Vice Chairman of pressure group
(ii)	The result of LPAT made me speechless (31/05/12)	OpEd	HK Headline online news service 'Double Talk' column	±	Journalist/Media commentator
(iii)	Can teaching return to its former glory? (9/01/08)	OpEd	Ming Pao newspaper (in English)	–	Student
(iv)	English proficiency important no matter how you say it (3/11/07)	Letter	Unknown	±	Student (teacher trainee)
<i>Pre-revision articles</i>					
1	Motivation is the missing link (24/6/06)	Letter	SCMP	+	Teacher
2	Language test for teachers a failure (24/6/06)	OpEd	SCMP	–	Teacher
3	Answer to the question 'Do teachers suffer more stress than other professionals?' (19/6/06)	Letter	SCMP	–	Doctor?
4	Teaching English demands competence (17/6/06)	Letter	SCMP	±	Unknown
5	First the teacher, then the learning (1/6/06)	OpEd	SCMP Education Post	±	Director of the British Council HK (Academic)
6	Job fears for 1500 teachers (23/5/06)	News	HK Standard	–	Journalist
7	A reason to learn E is vital (response to May 25 letter?)	Letter	SCMP	+	Unknown
8	The real role of teachers is to teach	Letter	SCMP	–	Unknown

(continued)

(continued)

No.	Article title (date of publication)	Type	Source	Stance	Writer
9	Job fears for 1500 teachers: 10% failure to meet the English language benchmark 'an unhappy indicator of our system' (23/5/06)	OpEd	Sing Tao newspaper (in English)	±	Journalist
10	Consistent weak and strong areas (23/5/06)	OpEd	SCMP City section	±	Journalist
11	2000 language teachers fail to make the grade (23/5/06)	News	SCMP	+	Journalist
12	Leniency urged for teachers in language test crisis (27/2/06)	News	SCMP City section	-	Journalist
13	Preliminary deal struck on ways to cut teacher stress (17/1/06)	News	SCMP	±	Journalist
14	Education chief sorry over suicide remarks (11/1/06)	News	Sing Tao newspaper (in English)	+	Journalist
15	The A to Z of reforms that have educators hopping (11/1/06)	OpEd	SCMP	-	Education Journalist
16	Institutions link up to lift the standard of English teaching (14/12/06)	News	SCMP City section	-	Journalist
17	Secretary dismisses claims of possible teacher shortage	News	SCMP City section	+	Journalist
18	Crisis looms as teachers fail crucial test	News	SCMP City section	-	Education Journalist
19	Training raises the bar for English teachers: programmes are on offer to improve teaching standards in line with language proficiency requirements laid down by the EMB (16/7/2005)	News	SCMP	+	Journalist
20	Overburdened teachers need support to perform better (18/6/05)	Letter	SCMP	-	Unknown
21	Experts say educators who fail to pass benchmarks exams must immerse themselves in the language to improve their skills (6/6/05)	News	SCMP	+	Journalist

(continued)

(continued)

No.	Article title (date of publication)	Type	Source	Stance	Writer
22	Poor English teachers must mind their language (30/5/05)	News	SCMP	+	Journalist
23	Language tuition in schools 'under threat from test' (28/5/05)	News	SCMP Education section	–	Journalist
24	Untitled (27/5/05)	Letter	HK Standard Metro section	±	Unknown
25	Just 29% of teachers get pass in English test (15/12/04)	News	SCMP City section	–	Journalist
26	Language teachers fall below mark (7/12/05)	News	HK Standard Student section	–	Journalist
27	Writing paper marked carefully under checks and balances (18/9/04)	Letter	SCMP Education section	+	Manager of LPATE
28	Teachers on the bench as they fall short of the mark (19/6/04)	Letter	SCMP Education section	+	Academic
29	Wide range of options considered for standards (19/6/04)	Letter	SCMP Education section	+	Manager of LPATE
30	Language education remains a top priority (19/6/04)	Letter	SCMP Education section	+	EDB?
31	Untitled (14/6/04)	Letter	SCMP Talkback	–	Teacher 'for some years'
32	Untitled (14/6/04)	Letter	SCMP Talkback	–	Unknown
33	Untitled (14/6/04)	Letter	SCMP Talkback	–	Teacher 'for a long time'
34	Untitled (12/6/04)	Letter	SCMP Talkback	–	Academic
35	Untitled (14/6/04)	Letter	SCMP Talkback	–	Teacher
36	Untitled (14/6/04)	Letter	SCMP Talkback	±	Academic
37	Untitled (14/6/04)	Letter	SCMP Talkback	+	Academic

(continued)

(continued)

No.	Article title (date of publication)	Type	Source	Stance	Writer
38	Untitled (14/6/04)	Letter	SCMP Talkback	±	Unknown
39	Untitled (14/6/04)	Letter	SCMP Talkback	+	Academic
40	Third of English teachers fail key test (10/6/04)	News	SCMP	–	Journalist
41	English teaching skills plummet (10/6/04)	News	SCMP City section	–	Journalist
42	Teachers in English struggle (10/6/04)	News	SCMP	–	Journalist
43	Language, power and testing; government hobgoblins stalk an ill thought out exam that wrecks teachers' morale every year (5/6/04)	OpEd	SCMP Education section	–	Academic
44	All eyes on the 'benchmark' (5/6/04)	News	SCMP Education	–	Journalist
45	LCQ13 Language proficiency requirement for teachers	Legco question response	Legislative Council records	±	EDB response to PTU questions
46	English degree launch (12/3/04)	News	SCMP	±	Journalist
47	Teachers fail to get higher marks (20/1/04)	News	SCMP	–	Journalist
48	Experts helped set teacher test (9/1/04)	Letter	SCMP	+	EDB
49	Dubious English test (31/12/03)	Letter	SCMP	–	Unknown
50	Educators demand tests' disclosure (20/12/03)	News	SCMP	–	Journalist
51	Teachers must mind language (8/12/03)	News	SCMP	±	Journalist

Key

SCMP South China morning post, OpEd opinion piece

Stance + positive, – negative, ± neutral

References

- Andrews, S. (2001). The language awareness of the L2 teacher: Its impact upon pedagogical practice. *Language Awareness*, 10(2 & 3), 73–90.
- Andrews, S. (2007). *Teacher language awareness*. Cambridge: Cambridge University Press.
- Australian Language and Literacy Council (1996). *Language teachers: The pivot of policy: The supply and quality of teachers of languages other than English*. Canberra: NBEET.
- Broadfoot, P., & Pollard, A. (2000). In A. Filer (Ed.), *Assessment: Social practice and social product* (pp. 11–26). London: Routledge.
- Brookhart, S., & Loadman, W. (1995). Perspectives on teacher assessment goals and their associated methods. In S. W. Soled (Ed.), *Assessment, testing and evaluation in teacher education* (pp. 9–39). Norwood, NJ: Ablex.
- Calhoun, C. (1992). Introduction. In C. Calhoun (Ed.), *Habermas and the public sphere* (pp. 1–48). Cambridge, MA: MIT Press.
- Coniam, D., & Falvey, P. (1999, June). *English language benchmark subcommittee pilot benchmark assessment report* (Report submitted to the Advisory Committee on Teacher Education and Qualifications).
- Diez, M. (2002). How will teacher education use assessments? An assessment scenario for the future. In R. Lissitz & W. Schafer (Eds.), *Assessment in educational reform: Both means and ends* (pp. 66–79). Boston: Allyn and Bacon.
- Dowson, C., Bodycott, P., Walker, A., & Coniam, D. (2000). Education reform in Hong Kong: Issues of consistency, connectedness and culture. *Education Policy Analysis Archives* 8(24). <http://epa.a.asu.edu/epaa/v8n24.html>, Accessed November, 2017.
- Drave, N. (1996). *The language of NVQs* (Unpublished Master's thesis). University of Birmingham, Birmingham, UK.
- Drave, N. (2006). The Language Proficiency Assessment for Teachers of English (LPATE) as an instrument of educational change. In *Proceedings of the 9th Academic Forum on English Language Testing in Asia* (pp. 18–40). Taipei, Taiwan: CEEC.
- Education Commission. (1995). *Education Commission report number 6*. Hong Kong: Government Printer.
- Fairclough, N. (1995). *Critical discourse analysis: The critical study of language*. London: Longman.
- Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.
- Filer, A. (Ed.). (2000). *Assessment: Social practice and social product*. London: Routledge.
- Fraser, N. (1993). Rethinking the public sphere: A contribution to the critique of actually existing democracy. In B. Robbins (Ed.), *The phantom public sphere* (pp. 1–32). Minneapolis: University of Minnesota Press.
- Fullan, M., & Stiegelbauer, S. (1991). *The new meaning of educational change*. New York: Teachers College Press.
- Giddens, A. (1990). *The consequences of modernity*. Palo Alto, CA: Stanford University Press.
- Glenwright, P. (2002). Language proficiency assessment for teachers: The effects of benchmarking on writing assessment in Hong Kong schools. *Assessing Writing*, 8, 84–109.
- Guthrie, G. (2011). Teacher resistance to change. In G. Guthrie (Ed.), *The progressive education fallacy in developing countries* (pp. 61–76). Netherlands: Springer.
- Habermas, J. (1966). *Between facts and norms* (W. Rehg, Trans.). Cambridge, MA: MIT Press.
- Habermas, J. (1991). *The structural transformation of the public sphere: An inquiry into a category of bourgeois society* (T. Burger, Trans.). Cambridge, MA: MIT Press.
- Habermas, J. (1993). The public sphere. In C. Mukerji & M. Schudson (Eds.), *Rethinking popular culture* (pp. 398–404). Berkeley, CA: University of California Press.
- Herman, E., & Chomsky, N. (1988). *Manufacturing consent*. New York: Pantheon Books.
- Hirai, M. (2002, September). Correlations between active skill and passive skill test scores. *JALT Testing & Evaluation SIG Newsletter*, 6(3), 2–8.

- Köksal, H. (1995). *Reducing teacher resistance to innovations*. Paper presented at the 6th IFIP World Congress, Aston University, Birmingham, UK. <https://linc.mit.edu/linc2013/proceedings/Session10/Session10Köksal.pdf>, Accessed November, 2017.
- Moore, A. (2004). *The good teacher: Dominant discourses in teaching and teacher education*. London: Routledge Falmer.
- Poedjosoedarmo, G., & Hsui, V. (1996). *Proficiency tests as predictors of student performance*. <http://www.aare.edu.au/96pap/poedg96435.txt>, Accessed November, 2017.
- Priestly, M., & Higham, J. (2002). New Zealand's curriculum and assessment revolution. *Education Line*.
- Reinemann, C., Stanyer, J., Scherr, S., & Legnante, G. (2011). Hard and soft news: A review of concepts, operationalizations and key findings. *Journalism*, 13(2), 221–239.
- Rich, C., Barcikowski, R., & Boyd, E. (1995). The assessment of preservice teachers. In S. W. Soled (Ed.), *Assessment, testing and evaluation in teacher education* (pp. 83–131). Norwood, NJ: Ablex.
- Richards, J., Gallo, P., & Renandya, W. (2001). Exploring teachers' beliefs and the processes of change. *PAC Journal*, 1(1), 1–36.
- TESOL. (2010). *Standards for the recognition of initial TESOL programs in P–12 ESL teacher education*. [https://www.tesol.org/docs/default-source/advocacy/the-revised-tesol-ncate-standards-for-the-recognition-of-initial-tesol-programs-in-p-12-esl-teacher-education-\(2010-pdf\).pdf?sfvrsn=4](https://www.tesol.org/docs/default-source/advocacy/the-revised-tesol-ncate-standards-for-the-recognition-of-initial-tesol-programs-in-p-12-esl-teacher-education-(2010-pdf).pdf?sfvrsn=4), Accessed December, 2017.
- Urmston, A. (2002). *The language proficiency assessment for teachers of English in Hong Kong: The first steps towards a fully qualified profession*. Paper presented at the 5th Academic Forum on English Language Testing in Asia, Tokyo, Japan.
- Urmston, A. (2003). *Do they do what they say they do? Comparison between assessor feedback and performance in a high-stakes English speaking test*. Paper presented at the 6th Academic Forum on English Language Testing in Asia, Seoul, Korea.
- Watson, J. (1998). *Media communication: An introduction to theory and process*. London: Macmillan.
- Zimmerman, J. (2006). Why some teachers resist change and what principals can do about it. *NASSP Bulletin*, 90, 238.

Neil Drape is a Senior Manager in the Assessment Development Division, HKEAA. He coordinates the development and marking of the HKDSE English Language and HKDSE Literature in English examinations, as well as the Language Proficiency Assessment for Teachers.

Part IV How Far Have Teacher Language Standards Improved Since the Inception of the LPATE in 2001?

David Coniam, Peter Falvey and Yangyu Xiao

This part describes the final element of the book's description of the series of steps, processes and products of the development of the LPATE in Hong Kong. Parts I–III covered the inception and development of the LPATE; an account of training and development programmes funded by the HKSAR Government for teachers who wished to improve their English language proficiency; the first revision of the LPATE in 2006; and research into the LPATE from an HKEAA perspective. Part IV now moves the narrative on 15 years in order to provide a stakeholder perspective of the impact of the LPATE on the teaching profession. The section consists of descriptions of studies which investigated the impact of the LPATE 12 years after its implementation in 2002—from the perspective of end users and related stakeholders. It consists of two chapters reporting a quantitative and a qualitative study.

Chapter 16

A Quantitative Investigation of Stakeholder Perceptions



David Coniam, Peter Falvey and Yangyu Xiao

Abstract This chapter reports on a quantitative investigation, conducted in 2015, of the perceptions of the end-users of the LPATE and related stakeholders after 15 years of LPATE administrations. The objectives of the study were to investigate, quantitatively, perceptions of the extent to which English teachers' English language standards may have improved since the introduction of the LPATE in 2000 and their perceptions of the impact of the LPATE [NOTE 1].

The Research Team

The research team consisted of the following:

D. Coniam (✉) · P. Falvey · Y. Xiao
Department of Curriculum and Instruction, Faculty of Education & Human Development, The
Education University of Hong Kong, 10 Lo Ping Road, Tai Po, Hong Kong
e-mail: coniam@eduhk.hk

P. Falvey
e-mail: falvey@eduhk.hk

Y. Xiao
e-mail: shirleyxiaoyy@gmail.com

Member	Position	Affiliation
David Coniam	Chair Professor and Principal Investigator	Department of Curriculum and Instruction, The Education University of Hong Kong
Peter Falvey	Adjunct Associate Professor	Department of Curriculum and Instruction, The Education University of Hong Kong
Alan Urmston	Assistant Professor	English Language Centre, the Hong Kong Polytechnic University
Neil Drave	Senior Manager, Assessment Development (English)	Hong Kong Examinations and Assessment Authority
Barley Mak	Associate College Head	United College, The Chinese University of Hong Kong
Yangyu Xiao	Senior Research Assistant	Department of Curriculum and Instruction, The Education University of Hong Kong

Survey

In this section, the design and preparation of the investigative survey used for the study is described. This is then followed by how, after piloting, the survey was administered online to invited teachers. An analysis of the data is then presented, and implications for stakeholders are discussed.

Data Collection Procedures

This section reports the data collection procedure for the quantitative study, including the questionnaire, which was administered via the Internet (and the *surveymonkey* online facility) from April to August 2015.

A major issue was to identify suitable target English language teachers. Ideally, responses would be received from teachers who had been in post for more than 10 years and also ideally from those who were now in leadership positions such as heads of department, vice principals, principals.

The first phase of the study included a questionnaire to be sent out to English language teachers around Hong Kong. The research team met in late 2014 and early 2015 to discuss issues related to questionnaire design, with the questionnaire worked on and revised both during and after the meetings via email. The first pilot questionnaire was trialled with a number of secondary English language teachers who proffered comments on the draft in early 2015. After moderating the questionnaire and making modifications, the survey instrument was then finalised.

Items were posed on a six-point Likert scale, with '1' indicating a positive response or agreement, and '6' a negative response or disagreement. A six-point scale was

Table 16.1 Survey—target participants

Target participants	Mode of contact	Number
Secondary school English teachers	Email invitations from the Hong Kong Examinations and Assessment Authority (HKEAA)	252
Primary and secondary school principals	Written handout/invitation via the Asia-Pacific Centre at the Hong Kong Institute of Education (HKIEd)	136
Former students of members of the research team	Telephone invitations	52
Primary school teachers on the HKIEd in-service ‘Primary Curriculum School Leaders’ (PCSL) programme commissioned by the HK EDB	Email invitations to the 2013–14 PSCL training programme cohort	45
Participants attending four-week weekend teacher training courses at HKIEd	Written handouts/invitation	200
	Total	685

deliberately chosen to prevent respondents sitting on the fence and not committing themselves to an opinion. Respondents were also asked to provide written comments on any aspect of the LPATE that they wished to. The questionnaire (see Appendix A ‘[English Language Proficiency Standards 2015](#)’) consisted of 33 items in four main sections. In Sect. 1, items 1–12 consisted of respondents’ personal and school details; Sect. 2 consisted of 2 items, items 13 and 14, providing details of the respondents’ passes on previous LPATE administrations; Sect. 3 consisted of items 15–18 and asked respondents for their views on the English language proficiency of English language teachers; Sect. 4, consisting initially of items 19–29 asked for respondents’ views of the ramifications of the LPATE. Finally, items 30–33 asked a general question about the effect of the LPATE and solicited respondents’ assistance in following up the questionnaire with a structured interview.

As the study ideally wished to obtain responses from long-serving teachers, teachers who were likely to have longer teaching experience were approached through a number of channels, as outlined in Table 16.1.

As can be seen from Table 16.1, a major channel to stakeholders was via the Hong Kong Examinations and Assessment Authority (HKEAA). Two hundred and fifty-two emails were sent out by the HKEAA to serving English teachers to obtain their agreement in late February 2015. Two hundred and sixteen teachers responded that they were willing to participate. The research team subsequently sent out the link to the online survey to those who had agreed to participate.

Table 16.2 Survey—response rates

	Number
Total number of questionnaires sent out	649
Total number of responses	289, including 272 online responses and 17 questionnaires returned through the post
Total number of completed responses	236
Response rate	44.5%

Other sources of tapping potentially worthwhile participants were as follows:

1. Serving teachers enrolled on the *Primary School Curriculum Leaders'* programme at the HKIEd (now re-named The Education University of Hong Kong) in mid-February 2015.
2. Primary and secondary school principals around Hong Kong who were approached by mail. Hard copy questionnaires with prepaid envelopes were sent to 136 principals in early May 2015. Seventeen completed questionnaires were returned. These hard copy responses were then manually input into *surveymonkey*.
3. The research team approached a number of potential respondents from former PGDE students of members of the research team in early June, inviting them to participate in the study.

As the questionnaire responses were anonymous, it was difficult to identify how many responses were collected from each channel, except for the hard copies returned. Response rates are presented in Table 16.2.

By mid-August 2015, 289 questionnaire responses had been collected. Of these, 236 comprised complete responses, 49 incomplete responses and 4 disqualified responses.

The response rate, as reported above, was 44.5%. Baruch (1999) states that the average response rate to paper surveys is in the region of 55% (p. 429). Nulty (2008), in a summary of studies of both online and paper surveys, reported that paper administration of a questionnaire results in a higher response rate. Given that the current study is something of a mixed-medium survey (although essentially online), the response rate of 44.5% may be viewed as acceptable.

Data Analysis

An analysis of the survey data will now be presented.

First, the robustness of the instrument is investigated. This may be gauged through reliability analysis and factor analysis. A presentation of key descriptives is then made—followed by an exploration of the inferential data.

Reliability

The first step in assessing the reliability of a questionnaire involves using the Cronbach alpha statistic. The analysis of the 14 attitudinal items on the questionnaire via Cronbach's alpha returned a figure of 0.799. The analysis suggested that item 17 ("Further improvement required") was somewhat problematic; removing this item improved the alpha to 0.825. Given that a level of 0.8 is generally recommended as desirable in a questionnaire (e.g. Dörnyei, 2003), this suggests that the questionnaire was generally well constructed.

Factor Analysis

An exploratory factor analysis using principal component analysis (PCA) with varimax rotation was conducted (working on the assumption that the underlying factors in the survey are related) to explore how the major constructs of the questionnaire were patterned and whether these fitted the two attitudinal sections that comprised the questionnaire, as laid out in Table 16.3.

In line with Kaiser's (1974) recommendations regarding sampling adequacy measures—the KMO (Kaiser-Meyer-Olkin) statistic—the figure of 0.807 indicated that the sample size was adequate for factor analysis. Bartlett's test of sphericity $\chi^2(91) = 1412.81, p < 0.001$, indicated that correlations between items were sufficiently large for PCA. An initial analysis was run to obtain eigenvalues for each component in the data. Four components had eigenvalues over Kaiser's criterion of 1 and, in combination, explained 67.83% of the variance.

Taking loadings above 0.4 as indicative of a cut-off point appropriate for interpretative purposes (see e.g. Stevens, 2002), four possible factors emerge in the component matrix. Table 16.4 elaborates.

As can be seen from Table 16.4, items 15–17 constituted one factor. Items 22 and 26 appeared to be in a factor of their own. Items 19–29 had three items which crossloaded onto two other factors. In general, however, the items appeared to fall into two factors, as per their design in the questionnaire. That is, items 15–18 probed fairly broadly views on English language teachers' English language proficiency. Items 19–29 probed issues about the impact of the test.

Table 16.3 Attitudinal sections of questionnaire

Section	Items
3: Views on the English language proficiency of English language teachers	15–17
4: Impact of the LPATE	19–29

Table 16.4 Component matrix

Components →	1	2	3	4
15. HK English teachers' proficiency acceptable to me		0.898		
16. HK English teachers' proficiency acceptable to most stakeholders		0.818		
17. Improvement in HK English teachers' proficiency required		0.452		
19. Minimum standard for English language teaching purposes	0.427		-0.415	
20. Introduction of LPATE important	0.849			
21. LPATE improved English proficiency	0.801			
22. Level required for LPATE Level 3 about right				-0.710
23. Preparing for LPATE improved own English proficiency	0.547		0.605	
24. LPATE helped get current teaching job			0.807	
25. All HK English teachers should take LPATE	0.812			
26. No exemptions or alternatives other than LPATE	0.649			0.448
27. EMI teachers should take an LPATE-type test	0.665			
28. Introduction of LPATE good for Hong Kong	0.882			
29. Potential HoDs chairs should attain LPATE Level 4	0.628			

Rasch Analysis

As mentioned in Sect. 1, Chap. 5, the principles of Rasch analysis differ from those of classical test analysis (CTA) in that Rasch enables different facets (the teacher questionnaire respondents and the items in the current instance) to be modelled on the same linear ruler, the same latent trait scale. Rasch data are usually provided in the form of logits, where 0 is the mid-point. With the current survey data, logit values below zero indicate disagreement and values above zero indicate agreement.

While the unit of the measurement scale is the logit, to aid interpretation of the results, 'Fair Averages' are provided in parentheses, to the left of the logit measures. Fair Averages (see Linacre, 1997, p. 550, for more details) are rating scale values converted from Rasch measures back to the original rating scale—the six-point scale in the current study. See Urmston, Chap. 13, this volume for a discussion of Fair Averages.

Since the result of the factor analysis indicated that the survey comprised two broad sections—and the fact that Rasch analysis should ideally be conducted on unidimensional data—Rasch analysis was conducted only with Sect. 4. Given that the survey consisted of two dimensions and that Sect. 3 comprised only three items (items 15–17), Sect. 3 was deemed too small for an individual Rasch analysis to be conducted. Rasch was therefore only performed with Sect. 4 (items 19–29), which consisted of 11 items. The software used to conduct the analysis was Winsteps (Linacre, 2006), which is based on the one-parameter Rasch model.

While the results from a Rasch analysis will be broadly comparable with a classical test analysis of the items, Rasch places both respondents and items on the same linear scale, permitting direct comparisons between both facets. To illustrate this, Fig. 16.1 presents the person-item map—a visual representation of how the two facets compare—where the persons (teacher respondents) are to the left and items to the right. Respondents with logit values above zero are positive in their responses and negative with logit values below zero. Likewise, items with logit values above zero are positively endorsed and items below zero are negatively endorsed. The mid-point of the scale—zero logits—equates with a Fair Average of 3.76 on the six-point scale, indicating that respondents were generally positive in their responses.

As can be seen from Fig. 16.1, teacher respondents were in a three-logit range, from -1.5 (disagree, FA 1.5) to +1.5 (agree, FA 5.0); one respondent was an outlier at +4.0 logits, strongly supporting all issues. Items were more closely clustered in a more narrow range of less than one logit, with only two items showing considerable divergence. These were item 19 (*There should be a minimum standard for English language teaching purposes*), on which there was agreement, and item 26 (*There should be no exemptions or alternatives other than the LPATE*), on which there was disagreement.

A crucial concept in Rasch is that of model fit, with ‘fit’ essentially being the difference between expected and observed scores. ‘Fit’ is defined differently by different researchers. Some researchers focus on the infit mean square figure. ‘Perfect fit’ is defined as 1.0, with acceptable practical limits of fit stated as 0.5 for the lower limit and 1.5 for the upper limit (see Lunz and Stahl, 1990; Weigle, 1998 for a discussion of limits of fit). Table 16.5 presents the picture of item fit.

Table 16.5 Item fit statistics

Item	Measure	S.E.	Infit mean square
19. Minimum standard for English language teaching purposes	-1.59	0.11	0.85
20. Introduction of LPATE important	+0.13	0.08	0.58
21. LPATE improved English proficiency	+0.41	0.08	0.54
22. Level required for LPATE Level 3 about right	-0.57	0.09	0.96
23. Preparing for LPATE improved own English proficiency	+0.30	0.09	1.03
24. LPATE helped get current teaching job	+0.40	0.09	1.51
25. All HK English teachers should take LPATE	+0.34	0.08	1.01
26. No exemptions or alternatives other than LPATE	+1.02	0.08	1.27
27. EMI teachers should take an LPATE-type test	-0.18	0.08	1.25
28. Introduction of LPATE good for Hong Kong	+0.10	0.08	0.43
29. Potential HoDs should attain LPATE Level 4	-0.36	0.09	1.25
Mean	0.00	0.09	0.97
S.D.	0.65	0.01	0.33

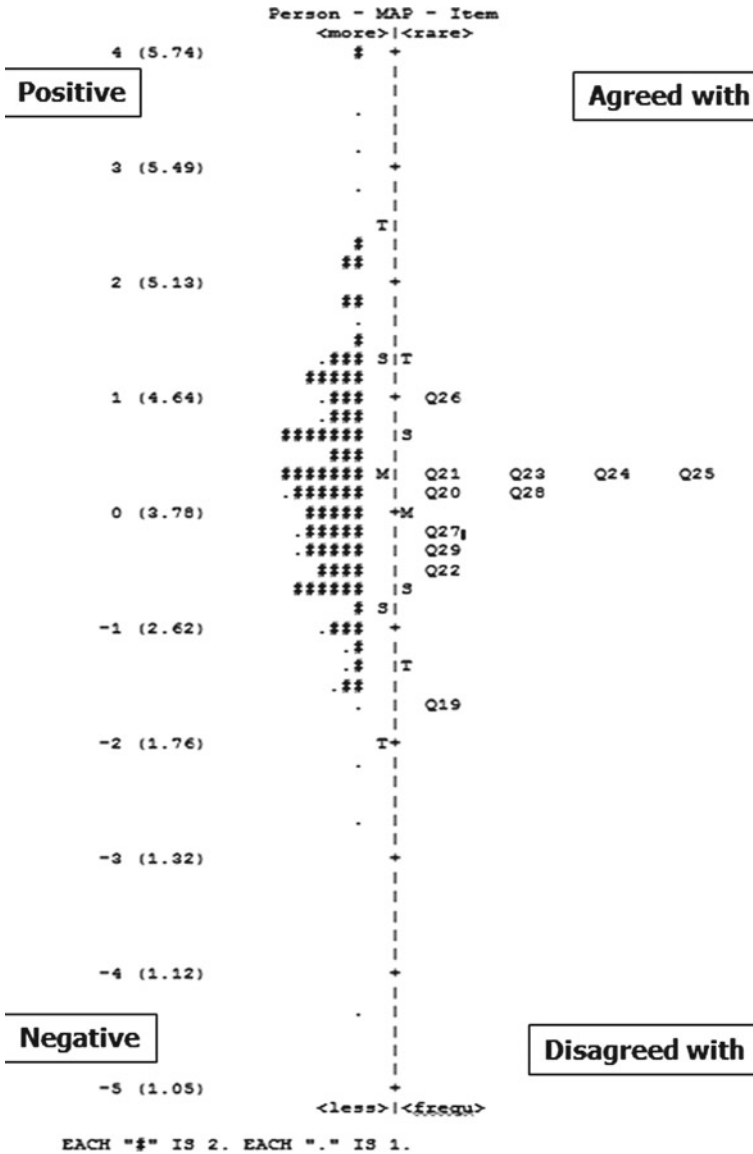


Fig. 16.1 Person-item map

As can be seen from Table 16.5, all items showed acceptable model fit—the exception being item 24, which slightly misfitted with an infit mean square just above 1.5. This can possibly be explained on the grounds that whereas the other 10 items were probing issues related to what had been done or what should be done in terms of policy, item 24 was asking a more personal, more concrete, question of

whether the LPATE helped teachers get their current job. Consequently, if they had been in post for a long time or were exempt, this question may not have had much relevance for them.

Descriptive Statistics

In the analysis below, markers' background details are first presented, with a subsequent examination of the questionnaires—first individually, and then contrastively. The Hong Kong Education Bureau (EDB) used to produce teacher surveys, usually at 10-year intervals. Against this backdrop, the large-scale survey that Falvey and Coniam (1997) conducted to gauge teacher reactions to the LPATE initiative in 1997 drew on the extensive Teacher Survey 1994 (Education Department Statistics Section, 1995) for comparative purposes. In this study, therefore, comparisons between the demographic variables in the current sample and the broader Hong Kong picture are made wherever possible. The latest available teacher survey, however, dates back to 2000 (see Census and Statistics Department, 2001). Consequently, while this is used where possible, comparisons are augmented by publicly available data from the Hong Kong Education Bureau (EDB) and other educational bodies in Hong Kong.

Of the valid responses, 151 (65.7%) were from English language teachers, 63 (27.4%) from heads of department and 16 (7.0%) from principals. Given the published figure of 9175 English language teachers and 934 primary and secondary schools in Hong Kong (Census and Statistics Department, 2001), the 10:1 teacher-to-principal ratio appears quite comparable. Many secondary schools in Hong Kong have both a lower (Years 7–9) and an upper (Years 10–12) form head of department for English; this amounts approximately to almost 2000 heads of department (934×2), and equates to a rough 4.5:1 head of department-to-teacher ratio. The head of department-to-teacher ratio in the current study is closer to 3:1. This figure is not wholly surprising given that part of the sample is comprised of former student teachers, known to the research team, a considerable number of whom have moved up in the hierarchy to management positions. Given these factors, the teacher to head of department to principal ratio in the sample is not too far off the actual ratio in Hong Kong schools.

Demographics

Before moving to an analysis of the data, it should be noted that two questions not asked in the survey concerned gender and age. Firstly, gender was not asked because the majority of teachers in school are female: Forrester (1997) reports a female–male English language teacher ratio of 8:1 in Hong Kong's primary and secondary schools; see also Census and Statistics Department (2001). The age variable was not included because age corresponds very closely with experience (see Falvey and Coniam, 1997), and experience was a factor which was recorded and analysed in the current study.

Concerning school type, 48 respondents (20.9%) were working in primary schools while 178 (77.4%) were working in secondary schools. This does not match the general Hong Kong picture of 54.9%: 45.1% primary to secondary schools ratio (Census and Statistics Department, 2001), but is understandable, given that the HKEAA's initial approach was to secondary teachers.

Of these, 183 (79.6%) were in aided schools, 18 (7.8%) were in government schools, 24 (10.4%) in Direct Subsidy Scheme (DSS) schools and 3 (1.3%) in private schools. This fits quite closely with the Hong Kong picture (see Committee on Home-School Co-operation, 2014).

Concerning school ability band, 63 respondents (27.4%) were teaching in a high-ability Band 1 school, 67 (29.1%) in a mid-ability Band 2 school, and 47 (20.4%) in a low-ability Band 3 school. This is a rough fit with the three ability bands that each covers 33% of the school population.

Regarding the medium of instruction, 144 (62.6%) were in Chinese as a Medium of Instruction (CMI) schools and 84 (36.5%) were teaching in English as a Medium of Instruction (EMI) schools. With a CMI-EMI split in Hong Kong of 73:27 (see Association of English Medium Secondary Schools, 2014), the sample does not diverge too far from the broader Hong Kong picture.

Concerning teaching experience, 4 (1.7%) had been teaching for less than 2 years, 21 (9.1%) had been teaching for between 2 and 5 years, 67 (29.1%) between 6 and 10 years, 43 (18.7%) between 11 and 15 years and 93 (40.4%) had been teaching for more than 16 years. This spread is difficult to compare with the wider general picture, given that the HKEAA would have been targeting experienced teacher-markers.

Hundred and ninety-three (83.9%) of respondents had a relevant degree (English language, English literature, Communication, Translation, Linguistics). This figure is also hard to compare, given that it has been rising consistently over the years (see Coniam and Falvey, 1999) since the inception of the LPATE in 2001, with the introduction of more degree courses, and a government policy of a degree qualification being a requirement for entry to the teaching profession.

As can be seen from the comparison of the demographic data with the broader Hong Kong picture, while there is not a perfect fit between the two sets of data, the survey sample in many cases is broadly comparable to the bigger Hong Kong picture. This suggests that the sample as it stands may provide quite a reliable anchor against the wider Hong Kong English language teacher depiction.

On the issue of how many times respondents had taken the LPATE, 84 (36.5%) had taken it once, 50 (21.7%) twice, 31 (13.5%) three times, and 11 (4.8%) more than 3 times.

On the question of the highest LPATE score obtained, only 123 responded—understandable since many with a relevant degree were exempt and did not need to take the LPATE [Note 1]. Table 16.6 presents the scores of those who responded.

As can be seen from Table 16.6, of the 123 respondents who had taken the LPATE, only a very small number had only taken it once. Over 50% had taken it three or four times. This may well be a result of the effect commented on by Coniam and Falvey (2003) regarding the manner in which a pass is calculated. In a given test component where band scales and descriptors are used to rate candidates (Speaking,

Table 16.6 Highest LPATE score obtained

	Frequency	Per cent (%)
Not taken	107	46.5
2	1	0.4
3	49	21.3
4	68	29.6
5	5	2.2

Writing, CLA), candidates must pass (achieve Level ‘3’ in effect) on *all* scales if they are to pass that component overall. This, Coniam and Falvey argue, results in a comparatively high ‘failure’ rate, with candidates needing to retake the test more than once.

Attitudinal Items

This section is in two parts. First, items which diverged greatly from the mid-point of 3.5 are discussed. Bradshaw (1990) (see also Coniam, 2013) proposes the term ‘consumer validity’, whereby a mean score considerably above (or below) the mean indicates strong acceptance of the proposition; i.e., that respondents wholly accept the argument. In the current study, a ‘6’ indicated a positive and ‘1’ a negative response; for convenience sake in the current dataset (see Table 16.7), this is taken as strong positive responses (bolded) being above ‘4’ or strong negative responses (italicised) being below ‘3’.

There was only one item with a low mean (below 3)—item 24. Item 24 (*LPATE helped me get my job*) had a figure of 2.94. While the figure suggests that the LPATE is not that critical in teachers finding an English language teaching position, the high standard deviation of the item (2.04) indicated a wide range of agreement.

Five items had means above 4 indicating strong agreement:

Item 15 (*HK English teachers’ proficiency is acceptable to me*) indicated that respondents felt that English language standards were generally acceptable. This picture was, however, slightly offset by item 17, which indicated that there was still a strong feeling that further improvement was nevertheless required. On item 19 (*There should be a minimum standard*), there was very strong agreement—at 5.01—that a minimum standard was necessary, essentially endorsing the government policy initiative to establish such standards.

On item 22 (*The level required for LPATE Level 3 is about right*), respondents believed that the minimum standard level for the LPATE of Level 3 had been pitched at about the right level.

On item 27 (*EMI teachers should take an LPATE-type test*), respondents felt that content subject teachers should face some form of minimum-standard LPATE-equivalent test—there being strong agreement at 4.04.

Table 16.7 Descriptive statistics

Survey item	<i>N</i>	Mean	S.D.
15. HK English teachers' proficiency acceptable to me	229	4.39	0.98
16. HK English teachers' proficiency acceptable to most stakeholders	229	3.83	0.58
17. Improvement in English teachers' proficiency required	229	4.31	1.09
19. Minimum standard for English language teaching purposes	230	5.01	0.91
20. Introduction of LPATE important	230	3.94	1.30
21. LPATE improved English proficiency	229	3.68	1.27
22. Level required for LPATE Level 3 about right	228	4.36	0.94
23. Preparing for LPATE improved own English proficiency	221	3.17	1.91
24. LPATE helped get current teaching job	222	2.94	2.04
25. All HK English teachers should take LPATE	229	3.77	1.68
26. No exemptions or alternatives other than LPATE	230	3.07	1.75
27. EMI teachers should take an LPATE-type test	230	4.04	1.37
28. Introduction of LPATE good for Hong Kong	230	3.92	1.29
29. Potential HoDs should attain LPATE Level 4	229	4.32	1.44

Finally, on item 29 (*Potential HoDs should attain LPATE Level 4*), respondents were unequivocal that head of departments should attain Level 4, with a mean of 4.32 recorded.

All these findings are indicative of a general feeling of approval for the LPATE, its assessment levels and its assessment criteria. These findings are very supportive of the HKSAR Government's decision to go ahead with the language benchmarking/LPATE initiative almost two decades ago.

Inferential Analysis

A chi square analysis of items where significant differences emerged will now be presented.

Hundred and forty chi square analyses were conducted—the 14 attitudinal items against 10 background demographic variables. In general, little significance emerged on the majority of the analyses, indicating that respondents were in agreement with items irrespective of backgrounds such as school type, whether they held a relevant degree, the school's medium of instruction or ability band. There were 17 instances of significance. Table 16.8a elaborates, presenting the data sorted by item, with attitudinal items in Column 1.

Table 16.8 a Attitudinal items—significant differences (sorted by item), b attitudinal items—significant differences (sorted by background variable); c interpretative commentary on crosstabs

Attitudinal items	Variables	Significance
24. LPATE helped get current teaching job	Primary or secondary school	$\chi^2(5) = 16.43, p = 0.006$
24. LPATE helped get current teaching job	Medium of instruction	$\chi^2(5) = 11.47, p = 0.043$
24. LPATE helped get current teaching job	Teaching experience	$\chi^2(20) = 34.98, p = 0.020$
24. LPATE helped get current teaching job	Times taken LPATE	$\chi^2(25) = 38.65, p = 0.040$
29. Potential HoDs should attain LPATE Level 4	School type	$\chi^2(15) = 31.88, p = 0.007$
29. Potential HoDs should attain LPATE Level 4	Highest level attained	$\chi^2(30) = 57.95, p = 0.002$
26. No exemptions or alternatives other than LPATE	Times taken LPATE	$\chi^2(25) = 40.70, p = 0.025$
26. No exemptions or alternatives other than LPATE	Highest level attained	$\chi^2(30) = 59.73, p = 0.001$
21. LPATE improved English proficiency	Primary or secondary school	$\chi^2(5) = 11.86, p = 0.037$
21. LPATE improved English proficiency	Times taken LPATE	$\chi^2(25) = 36.97, p = 0.058$
20. Introduction of LPATE important	Teaching experience	$\chi^2(20) = 33.11, p = 0.033$
20. Introduction of LPATE important	Ability band	$\chi^2(10) = 20.88, p = 0.022$
28. Introduction of LPATE good for Hong Kong	Times taken LPATE	$\chi^2(25) = 40.64, p = 0.025$
27. EMI teachers should take an LPATE-type test	Teaching experience	$\chi^2(20) = 33.16, p = 0.032$
25. All HK English teachers should take LPATE	Highest level attained	$\chi^2(30) = 61.17, p = 0.001$
17. Improvement in English teachers' proficiency required	Teaching experience	$\chi^2(20) = 34.89, p = 0.021$
23. Preparing for LPATE improved own proficiency	Primary or secondary school	$\chi^2(5) = 13.93, p = 0.016$

(continued)

Table 16.8 (continued)

(b)	Item	Variable	Commentary
	21. LPATE improved English proficiency	Times taken LPATE	$\chi^2(25) = 36.97, p = 0.058$
	24. LPATE helped get current teaching job	Times taken LPATE	$\chi^2(25) = 38.65, p = 0.040$
	26. No exemptions or alternatives other than LPATE	Times taken LPATE	$\chi^2(25) = 40.70, p = 0.025$
	28. Introduction of LPATE good for Hong Kong	Times taken LPATE	$\chi^2(25) = 40.64, p = 0.025$
	17. Improvement in English teachers' proficiency required	Teaching experience	$\chi^2(20) = 34.89, p = 0.021$
	20. Introduction of LPATE important	Teaching experience	$\chi^2(20) = 33.11, p = 0.033$
	24. LPATE helped get current teaching job	Teaching experience	$\chi^2(20) = 34.98, p = 0.020$
	27. EMI teachers should take an LPATE-type test	Teaching experience	$\chi^2(20) = 33.16, p = 0.032$
	25. All HK English teachers should take LPATE	Highest level attained	$\chi^2(30) = 61.17, p = 0.001$
	26. No exemptions or alternatives other than LPATE	Highest level attained	$\chi^2(30) = 59.73, p = 0.001$
	29. Potential HoDs should attain LPATE Level 4	Highest level attained	$\chi^2(30) = 57.95, p = 0.002$
	21. LPATE improved English proficiency	Primary or secondary school	$\chi^2(5) = 11.86, p = 0.037$
	23. Preparing for LPATE improved own English proficiency	Primary or secondary school	$\chi^2(5) = 13.93, p = 0.016$
	24. LPATE helped get current teaching job	Primary or secondary school	$\chi^2(5) = 16.43, p = 0.006$
	29. Potential HoDs should attain LPATE Level 4	School type	$\chi^2(15) = 31.88, p = 0.007$
	24. LPATE helped get current teaching job	Medium of instruction	$\chi^2(5) = 11.47, p = 0.043$
	20. Introduction of LPATE important	Ability band	$\chi^2(10) = 20.88, p = 0.022$

(continued)

Table 16.8 (continued)

(c)	Item	Variable	Commentary
	24. LPATE helped get current teaching job	Primary or secondary school	Primary school teachers were very strongly in agreement on this item
	24. LPATE helped get current teaching job	Medium of instruction	EMI teachers were more in agreement than EMI teachers, possibly because EMI principals tend to look for a first degree in English—the latter having conferred exemption from the LPATE
	24. LPATE helped get current teaching job	Teaching experience	A considerable number of teachers with the longest experience (16 years or more) disagreed on this item—probably because they were already in post
	24. LPATE helped get current teaching job	Times taken LPATE	Those who had taken the LPATE 2 or 3 times were more strongly in agreement
	29. Potential HoDs should attain LPATE Level 4	School type	While the numbers were small, government and DSS schools were strongly in agreement; aided schools ranged more widely across the agreement spectrum.
	29. Potential HoDs should attain LPATE Level 4	Highest level attained	The higher the level respondents had attained, the more in agreement they were
	26. No exemptions or alternatives other than LPATE	Times taken LPATE	The more times they had taken the LPATE, the more they tended to be in agreement
	26. No exemptions or alternatives other than LPATE	Highest overall	The higher the level they had attained, the more they were in agreement
	21. LPATE improved English proficiency	Primary or secondary school	Primary teachers were stronger endorsers; possibly because it is accepted that standards of English among primary teachers were lower 15 years ago when non-degree programmes were common
	21. LPATE improved English proficiency	Times taken LPATE	Those who had taken the LPATE 2 or 3 times were more strongly in agreement
	20. Introduction of LPATE important	Teaching experience	Those with the longest service (16 years or more) were the strongest endorsers, possibly because such teachers have perceived how standards have changed over time
	20. Introduction of LPATE important	Ability band	Teachers in Band 3 schools were not as much in support as those in Band 1 and 2 schools. English language issues are more of a concern in Band 1 and 2 than in Band 3 schools possibly because English is one of the keys to tertiary education—with few Band 3 students continuing on to tertiary education
	28. Introduction of LPATE good for Hong Kong	Times taken LPATE	Those who had attained higher levels tended to be more strongly in agreement
	27. EMI teachers should take an LPATE-type test	Teaching experience	The more teaching experience respondents had, the more they agreed
	25. All HK English teachers should take LPATE	Highest level attained	Teachers who had attained the highest level tended to be more strongly in agreement
	17. Improvement in English teachers' proficiency required	Teaching experience	Teachers with considerable teaching experience (11 years or more) were the strongest supporters, possibly because their longer time in the profession has permitted them to reflect on the need for improvement in teacher standards
	23. Preparing for LPATE improved own English proficiency	Primary or secondary school	Primary school teachers were more in agreement, again because of the earlier need to improve standards in primary school

Significance emerged on 10 of the attitudinal items (in Column 1), although against different background variables, and in general, with no discernible pattern. On item 24 (*The LPATE helped me get my current teaching job*), however, significance emerged against four background variables—possibly underscoring the importance the LPATE was having on recruitment after its introduction.

Table 16.8b presents a picture of the data sorted by background variable—in Column 2.

Focusing on Column 2, it can be seen that the number of times respondents had taken the LPATE was significant on four items, with the more times LPATE had been taken the stronger the agreement.

Length of teaching experience was another variable which emerged as significant on four items. This is in line with an earlier survey (see Falvey and Coniam, 1997) where there was a strong relationship between how long respondents had been teaching and how far they were in agreement with the notion that there should be agreed standards.

Whether teachers were teaching in primary or secondary school showed significance on three variables, with primary teachers more in agreement than secondary school teachers. This is possibly due to the fact that in the initial studies in the late 1990s (Coniam et al., 2000) primary school teachers tended to score lower on the LPATE by virtue of them not needing a degree to be a primary teacher (although they do now). Consequently, the general impression is that primary teachers' English language standards have risen considerably since the introduction of the LPATE: the introduction in Hong Kong teacher education institutions of compulsory degree programmes, subject to the overview of external examiners from within and without Hong Kong ensures acceptable levels of quality and standards.

Table 16.8c now presents an interpretation for each of the significant chi square crosstabs.

Candidature Over the Years

It should be noted that even though the majority of teachers are now granted exemptions because they have gained an undergraduate degree and/or postgraduate teacher qualifications, there is still a steady stream of candidates for the LPATE. It is noticeable that the pass rates for Reading have improved since 2010; that, as is pointed out in Chap. 8, Writing still poses the most problems for candidates; that Listening pass rates have also improved since 2011; that Speaking pass rates are better than they were in 2006; and that Classroom Language pass rates have maintained their high pass rates. Candidatures and pass rates are shown in Appendix B 'LPATE Results'.

LPATE entry fees for 2018 are HK\$2972 (approximately US\$380). The Reading, Writing and Speaking tests cost HK\$346 each; the Speaking test costs HK\$714 and the Classroom Language assessment component costs HK\$1220.

Conclusion

In this chapter, data collected from a questionnaire survey to gauge respondents' responses to a number of issues have been presented. The main findings were as follows: broadly, respondents were in agreement with the majority of the items asking about the need for a minimum standards test such as the LPATE, the Government policy initiative to establish minimum standards and the implementation of the LPATE. The general opinion was that English language teacher standards are now generally acceptable, although there was nonetheless a groundswell of opinion that further improvement was required. Respondents felt that the minimum standard level of the LPATE—set at Level 3—had been pitched at about the right level, and there was strong agreement that heads of department should attain Level 4.

Inferential analysis revealed that the number of times respondents had taken the LPATE to be significant on certain items, with the more times the LPATE had been taken the stronger the agreement; further, respondents who had attained higher levels in terms of LPATE results tended to be stronger endorsers of the test. Teaching experience was another variable which emerged as significant on certain items. Primary school teachers tended to be more in agreement about certain aspects of the impact of the LPATE policy than secondary school teachers, with teachers in Chinese medium of instruction schools also more in agreement than teachers in English medium of instruction schools. Finally, teachers in low-ability (Band 3) schools did not endorse the policy as strongly as did those in high and mid-ability (Band 1 and 2) schools.

In the next chapter, Chap. 17, the data gathered from the questionnaire survey are taken as the springboard to a qualitative in-depth investigation from a senior management perspective of the impact of the LPATE after having been administered for a period of fifteen years.

Notes

1. The research reported in this chapter was supported by the Hong Kong Research Grants Council (grant number 18401514).
2. The HK Education Bureau's Language Proficiency Requirements (LPR) for English language teachers can be found at <http://www.edb.gov.hk/en/teacher/qualification-training-development/qualification/language-proficiency-requirements/exemption.html>:

Applicants applying for exemption from the LPR (English Language) who are holding a relevant degree plus relevant teacher training will be granted full exemption from the LPR (English Language) and will be deemed to have reached Level 3 proficiency in the LPR (English Language).

Section 3: Views on the English language proficiency of English language teachers

Please indicate your level of agreement.

15. The English language proficiency of Hong Kong English language teachers is now at a level acceptable to me .	strongly disagree 1 2 3 4 5 6 strongly agree
16. The English language proficiency of Hong Kong English language teachers is now at a level acceptable to most stakeholders .	strongly disagree 1 2 3 4 5 6 strongly agree
17. Further improvement in the English language proficiency of Hong Kong English language teachers is required.	strongly disagree 1 2 3 4 5 6 strongly agree
18. Since I started teaching English language, English language proficiency of Hong Kong English language teachers has improved most in the following area (select one area only).	<input type="checkbox"/> Speaking skills <input type="checkbox"/> Writing skills <input type="checkbox"/> Listening skills <input type="checkbox"/> Reading skills <input type="checkbox"/> Interactive skills outside the classroom <input type="checkbox"/> Other (please specify) _____

Section 4: Impact of the LPATE

Please indicate your level of agreement.

19. There should be a minimum standard of English language proficiency for English language teaching purposes.	strongly disagree 1 2 3 4 5 6 strongly agree
20. The introduction of the LPATE was an important step in raising the English language proficiency of Hong Kong English language teachers.	strongly disagree 1 2 3 4 5 6 strongly agree
21. The LPATE has improved the English language proficiency of Hong Kong English language teachers.	strongly disagree 1 2 3 4 5 6 strongly agree
22. The level of proficiency required to attain Level 3 in the LPATE is about right.	strongly disagree 1 2 3 4 5 6 strongly agree
23. Preparing for the LPATE improved my own English language proficiency.	strongly disagree 1 2 3 4 5 6 strongly agree N/A
24. The LPATE qualification helped me get my current teaching job.	strongly disagree 1 2 3 4 5 6 strongly agree N/A
25. All Hong Kong English language teachers should be required to take the LPATE.	strongly disagree 1 2 3 4 5 6 strongly agree
26. There should be no exemptions or alternative ways of certifying Hong Kong English language teachers' English language proficiency other than the LPATE.	strongly disagree 1 2 3 4 5 6 strongly agree
27. Teachers of content subjects in EMI schools should be required to take an LPATE-type examination to certify their English language proficiency.	strongly disagree 1 2 3 4 5 6 strongly agree
28. In general, the introduction of the LPATE has been good for Hong Kong's education system.	strongly disagree 1 2 3 4 5 6 strongly agree
29. Hong Kong English language teachers wishing to become panel chairs should be required to attain Level 4 in the LPATE.	strongly disagree 1 2 3 4 5 6 strongly agree

- 31. Would you be available for a short follow-up interview? YES/NO
- 32. If your answer to Question 31 is Yes, please enter your contact details.

Your information will be kept in the strictest confidence and will only be used for contact purposes.

- 33. Please add any further comments about the LPATE and/or English language proficiency below.

Appendix B: LPATE Results

Year	Candidature	Reading (%)	Writing (%)	Listening (%)	Speaking (%)	CLA (%)
2001	396	86	33	68	51	89
2002	708	55	29	39	58	91
2003	1968	63	41	72	45	89
2004	2177	71	40	49	47	88
2004	1494	66	28	71	43	90
2005	1115	71	41	62	45	89
2005	1445	59	30	64	39	93
2006	953	86	46	74	37	93
2007	1836	79	40	80	48	93
2008	1285	83	42	72	62	95
2009	1298	80	46	70	51	97
2010	2058	66	43	72	44	94
2011	1867	89	37	83	50	96
2012	1826	88	37	83	50	95
2013	1739	89	45	78	52	98
2014	1631	84	53	83	52	98
2015	1625	88	61	86	55	97
2016	1524	87	50	85	55	97
2017	1471	85	39	83	56	97

Source http://www.edb.gov.hk/en/teacher/qualification-training-development/qualification/language-proficiency-requirement/lpat/lpat_assessment_reports.html

References

- Association of English Medium Secondary Schools. (2014). *Member list*. Available at <http://www.emi.edu.hk/>.
- Baruch, Y. (1999). Response rates in academic studies - a comparative analysis. *Human Relations*, 52, 421–434.
- Bradshaw, J. (1990). Test-takers' reactions to a placement test. *Language Testing*, 7(1), 13–30.
- Census and Statistics Department. (2001). *Major findings of the teacher survey*. Hong Kong Monthly Digest of Statistics, June 2001. Available at 2000 <http://www.censtatd.gov.hk/hkstat/sub/sp370.jsp?productCode=FA100198>.
- Committee on Home-School Co-operation. (2014, December). *Secondary school profiles 2014–2015*. Available at <http://www.chsc.hk/ssp2014/eng/index.php>.
- Coniam, D., & Falvey, P. (1999). Setting standards for teachers of English in Hong Kong—The teachers' perspective. *Curriculum Forum*, 8(2), 1–27.

- Coniam, D., Falvey, P., Bodycott, P., Crew, V., & Sze, M. M. P. (2000). *Establishing English language benchmarks for primary teachers of English language*. Hong Kong: Advisory Committee on Teacher Education and Qualifications.
- Coniam, D., & Falvey, P. (2003). Benchmarking the benchmark: Assessing the fit of a new test with its target population of teachers of English in Hong Kong. *Hong Kong Journal of Applied Linguistics*, 8(1), 1–15.
- Coniam, D. (2013). The increasing acceptance of onscreen marking—the ‘tablet computer’ effect. *Journal of Educational Technology & Society*, 16(3), 119–129.
- Dörnyei, Z. (2003). *Questionnaires in second language research: Construction, administration, and processing*. Mahwah, New Jersey: Lawrence Erlbaum.
- Education Department Statistics Section. (1995). *Teacher survey 1994*. Hong Kong Education Department: Government Printer.
- Falvey, P., & Coniam, D. (1997). Introducing English language benchmarks for Hong Kong teachers: A preliminary overview. *Curriculum Forum*, 6(2), 16–35.
- Forrester, V. (1997). The challenge of gender-bias reform: A case study of teacher trainees in Hong Kong. *Asian Journal of English Language Teaching*, 7, 113–119, <http://www.cuhk.edu.hk/ajelt/vol7/rep2.htm>.
- Kaiser, H. (1974). An index of factorial simplicity. *Psychometrika*, 39, 31–36.
- Linacre, J. M. (1997). *A user's guide to WINSTEPS/MINISTEP: Rasch-model computer programs*. Chicago, IL: Winsteps.com.
- Linacre, J. M. (2006). *Winsteps: Rasch measurement computer program*. Chicago, IL: MESA Press.
- Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Profession*, 13, 425–444.
- Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment & Evaluation in Higher Education*, 33(3), 301–314.
- Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Hillsdale, NJ: Erlbaum.
- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 263–287.

Chapter 17

A Qualitative Interpretation of the Impact of the LPATE on Key Stakeholders



David Coniam, Peter Falvey and Yangyu Xiao

Abstract Chapter 16 described the collection and analysis of the quantitative data in the research project funded by the HKSAR Government. This chapter describes the collection, analysis and findings of the data emanating from a complementary qualitative study. The chapter begins with a description of the data collection and its analysis, after which it then describes each of the major areas that emerged from the data analysis [Note 1].

Data Collection and Analysis

The initial data were collected through the survey questionnaire in mid-2015. This data was then supplemented by the collection of qualitative data from in-depth interviews with respondents.

Qualitative Interviews

The interviews were conducted in late 2015. As mentioned in Chap. 16, 236 complete responses to the survey were received through the online *SurveyMonkey* instrument by the end of August 2015. Of these 236, 57 respondents indicated that they were willing to participate in follow-up interviews. From this set, the research team then identified teachers who were English language heads of department or who had

D. Coniam · P. Falvey · Y. Xiao (✉)
Department of Curriculum and Instruction, Faculty of Education & Human Development,
The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, Hong Kong
e-mail: shirleyxiaoyy@gmail.com

D. Coniam
e-mail: coniam@eduhk.hk

P. Falvey
e-mail: falvey@eduhk.hk

Table 17.1 Interview participants

Post	Name	Years' teaching experience
Primary English language teachers (E)	Edwin (NET)	<2
	Emily	>15
	Eunice	>15
Primary English language head of department (H)	Helena	6–10
Secondary English language heads of department (S)	Samuel	>15
	Scott	>15
	Samantha	11–15
	Sophie	>15
	Stella	11–15
	Sara	>15
	Sabrina	>15
	Suzanna	>15
	Sandra	>15
	Sadie	>15
	Sylvia	>15
	Selena	11–15
Secondary English language teachers (T)	Tony	6–10
	Terri	2–5
	Tania	11–15
	Trudy	>15
Vice principal/principal (P)	Peter	>15
	Penny	>15
	Pamela	>15
	Peony	>15

Legend NET 'native-speaking English teacher'

considerable English language teaching experience (preferably both) and approached them about an interview. It should be noted that one NET (native-speaking English teacher employed from overseas—see <http://www.edb.gov.hk/en/curriculum-development/resource-support/net/index.html>) was also included to obtain something of an international perspective on the LPATE.

Table 17.1 describes the interviewees. Names have been anonymised for purposes of privacy. To aid readability, pseudonyms given to participants begin, where possible, with the first letter of their name reflecting the nature of their post. Primary teachers' names begin with 'E'; secondary teachers' names with 'T'; secondary heads of departments' names with 'S'; and (vice) principals' names with 'P'.

Table 17.2 Interview participants' posts

Post	No.
Principal/vice principal	4
English language head of department	13
English language teacher	6
NET teacher	1

Table 17.3 Interview participants' notional school bands

Participants		No.	Total
Secondary school teachers	Band 1	8	20
	Band 2	6	
	Band 3	6	
Primary school teachers		4	4

As can be seen from Table 17.1, the interviewees exhibit a wealth of relevant background. Twenty-two of the 24 participants had both a relevant degree and a relevant teacher training qualification, with 20 having substantial (more than 10 years) teaching experience.

Table 17.2 presents a summary of the posts they occupy.

The sample shows a set of participants with the management background that the study had been hoping for. Of the 24 participants, 13 were English language heads of department, with four on the principal track.

Finally, to illustrate the representativeness of the sample—and to indicate that participants did not come solely from high-ability band schools—Table 17.3 lays out participants' school bands. (Official bandings for primary schools, it should be noted, do not exist in the way that secondary schools did; hence, banding data for primary schools is not applicable.)

The composition of the participants currently teaching in schools was 20 at secondary level and 4 from primary schools. The split across school ability bands was quite even.

Interview Procedures

This chapter recounted the analysis of the quantitative study, and this chapter examines the interview protocols that were developed to investigate interesting issues arising from the quantitative study. After some initial pilot interviews in mid-2015, the interviews proper began, as mentioned, in late 2015. Typically, for interviews, 'easy' questions began the interview process. Participants were asked about their own backgrounds and their schools. They were also asked to comment on whether they felt

that English teachers' language standards had changed as a result of the LPATE, as well as any challenges they had encountered or observed. The reliability and validity of the responses were strengthened by asking respondents to provide examples and clarification for their responses. By late 2015, 24 interviews had been conducted, 20 from secondary schools and 4 from primary schools. All interviews were conducted in English—the language in which both interviewer and interviewees were competent and comfortable. Interviews were conducted in the interviewees' schools, with each interview lasting from thirty minutes to an hour, depending on the depth of the interviewees' responses.

Interview Data Analysis

The data analysis took a grounded and iterative approach using NVivo 8 and went on at the same time as the data collection. After being transcribed, the transcripts were carefully checked by the research team. The process of analysis of the data followed the usual pattern for NVivo 8, with the research team going through each transcript and identifying main themes relevant to the research questions of the research project. Meaningful units in the transcripts were first coded into free nodes in NVivo and were then put together under 'umbrella' tree nodes which included different ideas. All codes, including free nodes, tree nodes and child nodes were then revised after analysis of subsequent transcripts. After the initial analysis of all transcripts, the NVivo *Query* function was used to search for key issues that were deemed worthy of further analysis. For example, in this study, the query for the term 'grammar' helped identify all issues relevant to the relationship between grammar and the LPATE test, since 'grammar' had emerged as a key theme in the interview transcripts. The query for the term 'LPATE exemption' and 'exemption' put together all the views about exemptions for the LPATE. Conducting queries in this manner was therefore considered to be an efficient way of ensuring that all necessary and relevant information for a specific theme was coded. After the initial process was completed, there was a further re-reading and refinement of the codes.

Thematic Areas Derived from the Qualitative Data Analysis

Figure 17.1 presents the ten major areas/themes that emerged from the data. These have been grouped together in order to provide a coherent route through the areas that were uncovered during the qualitative data analysis.

- | |
|---|
| <ol style="list-style-type: none"> 1. Perceptions of the LPATE as a response to a problem in the past 2. The LPATE and teaching as a profession 3. The LPATE and language proficiency 4. The Language Proficiency of Hong Kong English Teachers 5. The LPATE and subject-matter knowledge 6. Pedagogical knowledge and skills 7. The negative side of the LPATE 8. Limitations of the LPATE 9. Changes over 12 years 10. LPATE requirements of English language teachers in Hong Kong |
|---|

Fig. 17.1 Thematic areas

The LPATE—A Response to a Problem in the Past

Respondents believed that the LPATE had considerable value in its early phases because, in the light of few teachers of English being both professionally and content trained, the LPATE was designed to remedy that deficiency by bringing all teachers of English up to the LPATE benchmark. Peter corroborated this fact by stating that many teachers of English were not language trained. He pointed out that many teachers whose major was not English were drafted into teach English classes which they had not been trained for.

Before the introduction of the LPATE (i.e. in the 1980s and 1990s), teachers without a qualification or relevant training in English were commonly asked to teach English because schools were short of English language teachers. Tony cited a teacher of Economics who taught him English but whose English was so deficient that his students mocked him. Stella stated that senior forms were given to English teachers with a degree in English while those without a degree in English taught junior forms.

The introduction of the LPATE, therefore, was one step towards ensuring that all English teachers were qualified and well trained. Peter stated that the introduction of the LPATE was important at that time in that it attempted to guarantee that English language teachers had a definite, acceptable standard of language proficiency. In this light, the LPATE policy was perceived by many participants as well intentioned. The LPATE policy, as Samantha recalled, removed many ‘unqualified’ English teachers from the profession—in particular in primary schools where many English teachers had majored in Mathematics or Geography. The introduction of the LPATE helped with maintaining English language standards and ensured that English teachers were capable and proficient second language users (Trudy).

The LPATE was, however, considered to be somewhat less relevant nowadays, as the comments below illustrate:

I don't see any impact brought by the LPATE because we don't have that many teachers taking it now. In my school, I remember only one teacher took the LPATE when it first

started. And then for a gap of 8-9 years, nobody needed to take it. And only recently another young teacher took it (Suzanna).

I think at that time, there are quite a lot of teachers they did not study English in the university, but yet the school asked them to teach English... then it is just a purpose at that point. Way back to 2000, just to screen them. Just to answer this call from society, saying our teachers are not qualified, are not trained, something like that (Suzanna).

The basic question is “What issues is the LPATE now addressing”? At that time it addressed a particular, and a very clear issue and problem confronting Hong Kong society. (Peter)?

There were few places for undergraduate in English Language Education up to the late 1990s. However, as Peter pointed out that situation has now changed with provision of many more undergraduate places both in Hong Kong and overseas so that there were now sufficient trained English language teachers.

The LPATE and Teaching as a Profession

Interviewees were clear that the LPATE had raised awareness that being an effective English language teacher required more than merely good English. It also required good levels of subject-matter knowledge and pedagogical knowledge.

A more positive way to view the expectations of English language teachers was that the LPATE gave English teachers a form of professional recognition. The LPATE policy delivered a strong message that schools needed individuals who were qualified to be English teachers. English language teachers, as with other professions, needed to have ‘professional’ knowledge of the language, as the comments below illustrate:

We need a way to tell the public that all these teachers are qualified. I think getting an LPATE qualification is one way for the public to see that teachers are professional. It’s not the same as thinking that anyone who has a degree can teach the language subject (Sara).

I think generally in society there are still a lot of people who have no idea that English teachers actually have got a professional knowledge in training, in English language, which does not really mean higher proficiency (Sandra).

With regard to what is meant by ‘being professional’, Sara made the point that English teachers should be proficient in their subject-matter knowledge irrespective of whether they taught high- or low-ability students. Sandra stated that teachers need to have subject knowledge, such as syntax, lexis, as well as pedagogical knowledge in teaching English, so that they are able to explain issues and concepts in English. Stella proffered a more pedagogically oriented response that teachers should be able to explain concepts in detail in English and be ready to mark students’ writing in ways that are beneficial to students. Another viewpoint—expressed by Samuel—was that professional English teachers are expected to be functionally adequate regardless of the level of the school or of the students. The LPATE is a qualification providing evidence of a teacher’s level of English.

The LPATE and Language Proficiency

The Importance of Language Proficiency

Many of the respondents asserted the importance of language proficiency for language teachers including Samuel, a panel chair, who made the point that good standards of proficiency were vital notwithstanding the ability of the students being taught. He added that language teachers' proficiency should be well above the standard required to teach students.

In Samuel's words:

You're teaching English. This is your profession. I keep on telling my teachers that despite being English majors, you have to keep your English at a certain level. Although you're teaching very weak students, don't put down your English newspapers. I think this applies to many schools in Hong Kong.

Eunice, a primary school teacher, agreed with Samuel's sentiments:

English teachers have to use accurate grammar, accurate language and accurate classroom language, because students are learning all the time. You don't know what they learn in the daily time. You also need to use language appropriate to the students' level. In primary schools, students are not having much vocabulary; you have to use the correct word choice for them. You have to have this kind of knowledge (Eunice).

Helena, a primary school panel head, commented that English teachers are likely to teach the wrong thing or wrong structures during students' early experience of learning English; thus, she was concerned about teachers' English language proficiency at primary schools. Tony agreed that English teachers' language proficiency is highly important, without which the teaching quality may have been compromised.

The primary school NET teacher, Peter, agreed that even in primary schools, high teacher language standards were important. He stated that the higher English teachers standards were, the more comfortable and confident teachers were when talking to students in English. Adding to the debate on the use of the mother tongue in the English language classroom, Peter asserted that teachers who were not confident in their own English proficiency tended to switch somewhat too quickly to Cantonese once they detected a lack of understanding in their students' eyes over what was being said in English. Students would then never listen to the English of these teachers because they knew a Cantonese translation would soon be proffered.

Suzanna, who worked in a Band 1 school, noted that students in top schools already had good language proficiency and so had high expectations of their teachers. The comment was also made that able parents could easily ascertain the English standard of teachers when talking to them (Pamela).

The LPATE as a Benchmark Test

In view of the importance of good language proficiency, the LPATE was viewed by many respondents as a suitable benchmark level for maintaining English language standards:

I think, that is, that (the LPATE) has actually helped to improve and to ensure the standard of English majors in Hong Kong... I think it may be a good way of screening teachers, the proficiency of teachers. (Samuel).

I guess as I have said, it is a benchmark, so if he or she is an English major, but they haven't taken the LPATE, well, ... I don't know their standard. So the LPATE to me is this kind of standard in the territory (Helena).

I mean you set the benchmark there, that you have to either get exemption or take the course or you have to pass the LPATE exam. Then that would certainly raise the standard of English teachers' proficiency (Emily).

Tony further illustrated how the LPATE maintained English teacher standards.

I think the LPATE is a good initiative because it really raises the English standard of the English teacher in Hong Kong in general. When I was in my previous school, I heard that some of the previous English teachers could not teach English because they couldn't get the qualification.

One outcome of the LPATE requirement, according to Tony, was that positive discrimination now took place in hiring teachers. The LPATE qualification was not needed before a teacher could be hired.

Sabrina noted how the LPATE provided teachers with feedback on their weaknesses:

I guess because of these sub-papers, the teachers would be better informed of their strengths and weaknesses in a certain area. This is important (Sabrina).

Samantha commented on how the LPATE also provided teachers with the chance to practise their English and thereby make improvement:

Like 10 years ago, I was worried about the LPATE, and I actually took a course at the PolyU. And I think I now pay more attention to grammar items, and to error correction and have a better idea of how to explain errors using relevant metalanguage – say for example there should be a to-infinitive after a verb, things like that. So in that aspect, I think there is some kind of improvement in myself. I mean at least I'm more aware of the metalanguage that I should use to explain errors to students (Samantha).

Samuel felt that all teachers should attain the LPATE benchmark. He commented on the usefulness of the scales and descriptors in providing valuable feedback, especially to those who failed. He said that the LPATE was therefore a good indicator for teachers to know where they were, what proficiency level they were at and what they had achieved. Emily commented on how the role of identifying weaknesses in teachers' language abilities in such a way illustrated the quasi-formative nature of the LPATE. How improvements might be made was, however, another matter and is the subject of further discussion below.

The Language Proficiency of Hong Kong English Teachers

When asked to comment on the language proficiency of English teachers at their school, six interviewees expressed a high degree of confidence about the language proficiency of teachers in their schools.

For our school, we never really had a very big problem in recruiting proficient teachers. We don't really find much difference because like even my other colleagues who use English as a medium of instruction, their English is just as good as mine (Penny).

I think I am quite happy with their English language proficiency. Maybe our school is very demanding in choosing English teachers (Suzanna).

As I said, the standard of the English teachers here is really high. Some of them are non-Chinese – Australians and Canadians, etc. So when comparing this school with my previous school, the standards of the language teachers are much higher (Trudy).

The responses above show that in schools where teachers were considered to have high language proficiency, either the school had a tradition or a high enough reputation to attract qualified teachers. The LPATE consequently had little impact on such schools.

A similar view, although not as strong as the positive views expressed above, was that teachers generally had sufficient language proficiency to communicate effectively in schools. The NET teacher Peter observed how his colleagues could communicate with him perfectly in English. Tania similarly noted how all the English teachers in her school could function and communicate perfectly acceptably in English. In Edwin's school, everything in the English department was conducted in English and Edwin commented on how he was impressed both with the accuracy of the grammar and the range of vocabulary used by teachers, and the quality of the language used in department reports.

While generally happy with the English language standards in her school, Pamela commented how she felt that English language standards in Hong Kong were deteriorating. English majors seldom read literature, she said, unlike their counterparts twenty years ago, so these teachers had a weak background knowledge of the literature. Sadie stated that in her opinion, although the English proficiency of teachers had improved generally, the English language proficiency of English majors nowadays was weaker than that of English majors twenty years ago, probably because of the expansion in university recruitment and her perception of the dilution in the quality of undergraduates.

A considerable amount of variation in the language proficiency of English teachers across school levels was reported. Drawing from her own experiences of working in different schools, Emily commented that the language proficiency of the English teachers in her primary school was lower than that of her colleagues in her previous secondary school. Language proficiency was also noted as varying between younger teachers and more experienced teachers. Sophie found that, in her school, younger teachers were better at speaking while experienced teachers were better at writing.

The LPATE and Subject-Matter Knowledge

The change in attitudes of key stakeholders emerged as a significant theme. Principals, panel chairs and teachers now generally accepted that not everyone should be allowed to teach English and that not everyone could reach the required level in the LPATE. It was generally accepted that all teachers of English should be trained in English. Scott stated that he had noticed the change in administrators and panel chair attitudes as a direct result of the LPATE requirement. He stressed the need not only for adequate language proficiency but also professional training.

Without sufficient subject-matter knowledge, it was suggested, teachers may teach students' incorrect concepts and mislead students (Eunice). To underscore the importance of subject knowledge, both Peter and Sara backed up their argument with examples such as themselves not being able to teach subjects such as Mandarin, Japanese or Maths because they did not have the relevant subject knowledge.

Two particular aspects of English language were identified in the data. These were knowledge of grammar and metalanguage. Terri explained how studying for the LPATE had assisted her in explaining grammar to students. As English teachers need to explain grammatical errors to students, an awareness and knowledge of grammar were important, a point reiterated by Helena. Sadie commented on how younger teachers tended to have less explicit knowledge of grammatical structures and how she felt that the LPATE had drawn attention to the need for a good foundation in grammar.

Sara believed that by preparing and taking the LPATE, English language teachers would be able to improve their knowledge of grammar and to clarify grammatical concepts. Stella remarked how many teachers now appeared to be more aware of the academic aspects and technical terms of the English language. English teachers, according to Sara, needed to have a very strong foundation of grammar in order to be able to explain errors—which would surely have attendant benefits to students when they faced public exams.

In the light of a communicative approach to teaching language, whether it was still necessary to teach or test grammar was a contentious issue in some of the interviews. Samantha felt that even though the LPATE improved candidates' grammatical knowledge, she noted that a communicative approach to language teaching placed less emphasis on grammar than did other methods.

Given that grammar is somewhat soft-pedalled in the public examinations, respondents such as Samantha and Suzanna felt that a comprehensive knowledge of grammar was not required, asserting that the consultation of a reference grammar would suffice if students asked questions that they could not answer. Another viewpoint was that teachers did not need to explain grammar to students using terminology such as that tested in the LPATE. It would be sufficient if teachers were able to tell students what was wrong and how to use the language correctly; metalanguage was not necessary (Samantha). These responses would suggest some teachers still do not appreciate the need for language awareness and grammatical understanding by the teacher (see Andrews, 2003).

Other teachers such as Samuel disagreed with the attitude above. Samuel said:

You're talking about a test for the English teachers. Accuracy is important because you are teaching English. If you yourself cannot even get the rules right, how can you teach students? You say "Oh the way I'm using is the communicative approach". Honestly, does that mean you don't need grammar? Does that mean the English that you or your students use need not be grammatically correct? It makes no sense, right? (Samuel)

Samuel was worried that the importance of grammar would be further downplayed if English teachers and heads of department considered it not necessary that grammar be tested in the LPATE. Samuel believed that all English teachers needed to be strong in terms of the accuracy of their English. Even if the accepted methodology in Hong Kong revolves around some form of communicative approach to language teaching, teachers and students still need to use grammatically correct English and teachers still need to be grammatically aware. The interviewees who supported the importance of grammar were of the view that second language teachers, as English majors, needed to know how the language worked explicitly and to have good knowledge of the terminology, with one argument being that students in both senior forms and in certain prestigious primary schools (where English was the medium of instruction) expected teachers to explain or demonstrate grammatical issues clearly to their students (Helena).

Pedagogical Knowledge and Teaching Skills

The Classroom Language Assessment Component

The LPATE component most closely related to language teaching is Classroom Language Assessment (CLA). As mentioned in previous chapters, CLA assesses teachers' use of language in their own English language classrooms. The CLA component was considered to assess a particular aspect of language proficiency that cannot be assessed through other language proficiency tests (Sara) in that CLA is directly related to the practice of teaching (Tony).

I would say the examination somehow channeled the teachers' training. In the past, perhaps there was no speaking paper, no classroom language. I do think nowadays, because of the speaking exam and perhaps classroom language, the new teachers are more competent in their communication (Sophie).

Sophie's response above illustrates how the LPATE drew explicit attention to proficient classroom language—a key aspect of good language teaching. The English language proficiency of teachers in lower band schools, for example, was reported to have improved after the introduction of the LPATE. On this issue, Peony noted that many English teachers in lower band schools, who in the past spoke Cantonese in class, were now using more English in their classes. Sara mentioned that teachers who had successfully gained the LPATE qualification were more confident when

teaching, writing minutes of English panel meetings and discussing in English during department meetings.

Some interviewees mentioned limitations of the CLA component. They suggested that performance on the CLA may also be affected by the motivation and language abilities of students or the relationship between teachers and students. Tania felt that teachers might score poorly if students were unwilling to respond and there was little interaction in the classroom. Helena felt that as the CLA assessment is now generally examined by only one examiner, it may not accurately reflect a teacher's ability

Pedagogical Knowledge and Teaching Skills

In addition to adequate language proficiency, respondents mentioned the importance of pedagogical and teaching skills:

A key issue relating to whether a teacher can teach effectively is how well the teacher has mastered the pedagogical skills (Sara).

For teaching English, language proficiency is one thing, but whether you have passion, something to do with your teaching methodology, whether you truly care about your students or not, are also important. I think these factors matter more than language proficiency itself (Samantha).

More specifically, English teachers, according to Tania, need to know how to help students, in particular low-achieving students, by such means as moderating the pace of their teaching, rephrasing to facilitate understanding, using group discussion strategies, using pictures as examples. Such pedagogical skills could also be related to subject matter, for example guiding students towards thinking critically and developing a piece of argumentative writing (Sandra). It should be noted that many of these points—e.g., moderating the pace of teaching (see Tania's comment above)—form part of the aim of the CLA of 'demonstrating language competence in presenting to and interacting with students' and are assessed on *The Language of Presentation/Practice* and *Pronunciation, Stress and Intonation* scales.

Sylvia, as a head of department, stated that she would consider non-English major teachers with the LPATE qualification, as long as they had the personality, experience and methods that were likely to help students. Good English teachers were also expected to have a positive personality, as well as a good attitude:

Subject knowledge is important but it cannot make you a good teacher. I think it's the character. The character, personality, mentality and attitude of a teacher, although they come second after subject language, are very vital (Samuel).

When recruiting teachers, we look into their personality, whether they could build up a good relationship with students (Sylvia).

Samuel, a veteran English teacher with over 30 years' experience, regarded himself as an 'energetic and vibrant teacher', one who always tried to give students the feeling that although English was not easy, they should be willing to try.

Responses in this section revealed that stakeholders feel that language teachers need pedagogical knowledge and related skills to cope with the day-to-day classroom and that these are no less important than language proficiency and subject-matter knowledge.

The Negative Side of the LPATE

The introduction of the LPATE policy was generally viewed as being a correct policy decision for Hong Kong, as it helped with maintaining/ensuring English language standards for all English teachers. The LPATE policy, nonetheless, it was noted, caused a considerable amount of controversy—in particular in the first few years of its implementation.

Participants recalled how, after the LPATE was introduced, many teachers protested about having to take the test, especially after having been teaching for many years (Pamela) and despite the fact that many were English majors (Sylvia). As a consequence of the protests and other negative publicity, the HKSAR Government rethought the policy and backtracked on not granting exemptions. While this was generally seen as a positive move, the exemptions were, in some quarters, not considered to be totally fair. The range of degrees for exemption, according to Samuel, was not convincing, in that it was difficult to justify why exemption was given for some degrees and not to others. Teachers with an English degree from the former colleges of education, for example, were still required to take the LPATE (Samuel).

Older and experienced teachers felt humiliated about having to take a test in their mid-fifties if they wished to keep their jobs (Pamela). Samantha commented on this feeling of humiliation in the context of it being closely linked to not feeling trusted:

I would use the word ‘humiliating’. I find it humiliating. That means you don’t trust me as an English major. My English should be OK. At least I could communicate with my students and could discuss educational matters with my colleague, so I really felt very bad about the idea of being asked to take the LPATE at that time, because I think A-Level would be quite an accurate estimation (Samantha).

Teachers felt there was a lack of trust between themselves and the government, as well as between the government and the universities. While Samuel encouraged all teachers in his department to take the LPATE in order to showcase their language standard, Sylvia and Suzanna were strongly opposed to such a move, stating that requiring teachers to take the LPATE showed how heads of department did not trust their teachers.

The experience of taking the LPATE had also been a nerve-wracking experience for some, with candidates—such as those who had not taken examinations for a long time—feeling anxious, pressured and concerned about failing—a potentially very embarrassing situation (Peony, Sabrina). Scott suggested that the pressure appeared to be more intense for experienced than for new teachers, as new teachers were

already aware of the requirements before they had made the decision to be English teachers.

Limitations of the LPATE

The first limitation raised by interviewees related to a number of test quality issues. The LPATE was criticised for being too difficult and for certain components not being well designed—in particular the Writing and Listening Tests. Sandra, drawing on her experience of communicating with teachers in her department, commented how some candidates had failed the Writing Test simply through not having completed all the sections of the test paper. Sandra's own experience of taking the LPATE trial test (the PBAE in 1999) and the first LPATE test (in 2001) reinforced her contention that the assessment was demanding and too difficult. Sadie and Peter agreed that the speed of delivery of the listening component was too fast, with candidates not having sufficient time to write down the answers—in particular during the first few years of the administration of the LPATE, and before the test was revised in 2007—as Urmston describes in Chap. 15.

The second limitation related to the validity of the LPATE and the nature of the LPATE as a test. With the Classroom Language Assessment component (since 2010) being only assessed by one examiner in one single lesson, Helena doubted whether one classroom observation was sufficient for a sound judgement of a teacher's language proficiency to be made [Note 2]. On a slightly cynical note, Tania suggested that it was possible for students to 'assist' the teacher and pretend to be cooperative or responsive. While such a setup might help a teacher score high, it could give a false picture of the teacher's English standard.

Some argued that, as the LPATE was regarded merely as a test, it might not necessarily reflect the language proficiency of English teachers, as the excerpts below exemplify:

Limitations... I think just like any kind of assessment. We can only test the teachers' English level in a one-time basis. It may not truly reflect the teachers' ability in the school setting (Sara).

Maybe if you follow the right pattern you will achieve higher marks (Terri).

Understanding the examination skills required by the LPATE was considered to be important. The comparatively new English teacher Terri explained how she failed in her first sitting of the LPATE because she had not prepared. The second time around, however, after careful preparation and study of past papers, Terri obtained Level 4, indicating the place of practice in obtaining a high score. Samantha agreed that while preparing for the LPATE, she felt she had become more aware of the examination format and examination skills rather than the language itself.

Teachers who had failed the LPATE described themselves as not being smart exam-takers. Selena, for instance, failed the Speaking Test twice because, as she put

it, she had not tuned into the ‘test taking game’. Sandra commented on how passing the LPATE required certain exam skills:

The LPATE exam involves exam skills. If you want to get a pass, there are a lot of exam skills – something that English teachers are not necessarily very good at. So requiring them all to take the LPATE is not really very meaningful unless the exam is somehow changed into something more like a professional exam – like really just for English teachers. Otherwise I don’t see the incentive for English teachers to do it (Sandra).

Samantha recalled that she became more aware of the exam format rather than the language itself through preparing for the LPATE.

Recalling her experience of taking the LPATE the first time, Tania attributed her lack of success to failing to making eye contact with the examiners:

I still remember I failed the first time, in Speaking I got 2.5, I think it is not because of my English proficiency in speaking but because I didn’t make eye contact with the examiner. Because it’s my first time to take the exam without any preparation, I just read aloud the poem, and then the examiners failed me. So I know my weakness, eye contact, and then I tried, I attempt the second time, and then I passed the second time (Tania).

The third limitation accused the LPATE of being an assessment *of* learning, rather than an assessment *for* learning (Sophie), and therefore contributing less than it might to teachers’ professional growth. Fourth, having to take the LPATE could be stressful for teachers, an issue which was considered to have an effect on teachers’ performance (Sara). Sara noted how, teachers at her school had been very stressed over having to take the LPATE because they had fallen out of the habit of taking examinations. Compared with older teachers, newly graduated teachers felt less stress, as new graduates were more used to taking examinations. Sara backed up this point with her own experience of taking the LPATE:

When I took the LPATE oral-the group discussion, even if I’m a very experienced oral examiner, I found that difficult. You know what, oh my god, the younger ones would, I mean, would do it pretty aggressively. If you don’t take part in the discussion and then you would fail. I remember that experience, so one thing that affects the performance could be the experience in taking examinations or whether they take it continuously...constantly (Sara).

The final limitation made was that the LPATE is now open to be taken by members of the general public. It was suggested that this issue has resulted in a rather lower pass rate—although it should be noted that the LPATE (excluding the CLA) has always been open to members of the general public. Some interviewees commented that with the LPATE being open to all, the lower pass rate had been giving the public the wrong impression that is that English teachers in Hong Kong were poor at English (Sandra and Peter). Peter stated that as anyone can now take the LPATE, the LPATE was not fulfilling the function it was set up for.

The limitations of the LPATE mentioned above included the difficulty level of the LPATE, the issue of test validity, the LPATE as an assessment *of* learning rather than an assessment *for* learning, the stress a one-off test may cause and the potentially wide range of candidates who may now sit the test. The limitations articulated may shed light on further improvement in the LPATE.

Changes Over 15 Years

Regarding the issue of any perceived changes in English teachers' language proficiency since the introduction of the LPATE, a number of issues were raised.

Many interviewees pointed out that one major change that occurred because of the introduction of the LPATE was that both professionally and in English content knowledge, teachers had become better trained. As alluded to above, the English language head of department Samuel reflected on the fact that while six years ago there were non-English major graduates who had English teaching posts in his school, this was no longer be the case. Pamela, a school principal teacher stated that as recently as 14 years ago only two universities (the University of Hong Kong and the Chinese University of Hong Kong) in Hong Kong offered undergraduate degrees in English, so it was difficult to recruit qualified teachers. She pointed out that teachers who had degrees in other subjects (e.g. economics and history) were instructed to teach some English classes. It was noted, therefore, that since the introduction of the LPATE, increasing numbers of English teachers in the respondents' schools were now subject trained. Former English teachers without a degree in English had either switched to teaching their own subject or had taken degree courses in English language (e.g. Terri, Samantha). The introduction of the LPATE had made teachers increasingly aware that to be an English teacher, expertise and formal training in English were now expected (e.g. Sylvia, Stella, Pamela and Scott).

There was a knock-on effect, however, with the professional training which English teachers had been receiving being seen as contributing to the rise in English language standards generally. Comparing current English teachers' language standards with English standards when Tony himself was a student, Tony felt that the professional training teachers now received did make them better language teachers. The younger teachers had better proficiency because they had taken relevant degree courses and passed the LPATE.

A direct impact on English language proficiency was not obvious or dramatic, some felt, as Samantha's words exemplify:

I really don't think there's much difference in the standard of language after the implementation of the LPATE, because I do not think there is a strong correlation between the two things. Maybe the teachers are more aware of language errors when they do the language error correction paper. But I think even without the LPATE, my colleagues are quite good at English and they are able to mark students' errors professionally (Samantha).

Sophie's school had a long tradition of recruiting high-quality English teachers, and for the past 12 years, the school had insisted on only recruiting teachers with a degree in English. Sylvia believed that all teachers had different strengths and weaknesses, so the situation in her school had not really changed much over the past decade.

The LPATE Requirements of English Language Teachers

The Need to Take the LPATE

Participating teachers held different views regarding whether it was necessary to take the LPATE, with a number expressing reservations about the necessity of having to take the assessment. A major reason for not taking the LPATE, some suggested, was that an English major degree was sufficient. Tania stated:

Why do English teachers have to take the LPATE if they have already demonstrated their proficiency when they entered the university, or undergraduate studies? So why should we have to test them again? I cannot see the logic (Tania).

The requirement for completing a university English degree was considered by some interviewees to be sufficient to guarantee the quality of English teachers. Sabrina commented on how all her colleagues who held an English degree had a high level of proficiency in English. If all English majors had to take the LPATE, some argued that there was little point in studying for a degree in English.

A different point of view put forward by some other heads of department, however, was that considerable variation in language ability among English major graduates existed and thus taking the LPATE should remain a necessity. Samuel made the point that an English degree does not necessarily guarantee English language standards. Consequently, his school required all their English teachers to take the LPATE. Sophie outlined how, when interviewing job applicants—even some English major graduates from prestigious Hong Kong universities—were found to be wanting in terms of their standard of English; further, she commented on how some English major teachers in her school had only managed to obtain Level 2 in the Speaking Test. In view of such variation, Helena suggested that the LPATE continue to be the standard in the territory—without which it would be difficult to compare the English language standard of English major graduates.

The second reason for not taking the LPATE, however, was that there were various ways in which English teachers' language proficiency might be certified, with IELTS, for example, being mentioned in this regard. Selena said that she would be confident about an applicant's language proficiency if they had an IELTS score of 7 or above. Samantha described the LPATE as somewhat repetitive in nature, as the current public examination system should be able to give an indication of candidates' English language standards. IELTS was now being taken by many Hong Kong university students as a form of graduation exit test [Note 3], and consequently, it should be possible to gauge graduates' language proficiency from their IELTS score. A counter argument to this suggestion was that the LPATE has a component assessing classroom language and a writing section focusing on explaining errors to students—neither of which appear in the alternative tests mentioned; further the LPATE also includes a relevant performance-based speaking component. Other reasons previously alluded to for not necessarily taking the LPATE were that passing the LPATE requires a considerable amount of examination skills and that taking

the LPATE causes unnecessary pressure and the administration of a test such as the LPATE involves a considerable resources, which might be made better use of for other educational purposes. Such objections appeared to somewhat miss the point, however, since IELTS is a high-stakes examination that also puts pressure on test-takers. Indeed, overall, it is important to remember that LPATE is a ‘special purpose’ form of assessment not a typical language proficiency test.

The Language Proficiency Level Requirements

Level 3 for English Language Teachers

The LPR set by Government was that English language teachers needed to achieve Level 3 in the LPATE. The need for English language heads of department to attain Level 4 came about through the schools’ determination rather than by legislation. Interview participants expressed their views towards this policy.

On the issue of the requirement of Level 3 for English language teachers, this level (‘3’) was generally considered to be an indication (a ‘guarantee’ in the words of some respondents—Sabrina, Helena) that English teachers met basic requirements—tallying with the response on the survey (Q.22: *The level of proficiency required to attain Level 3 in the LPATE is about right*—mean 4.29/6). Both the public and schools clearly expect teachers to have high English language standards, with Level 3 being the minimum acceptable standard if they are to be deemed qualified.

Four interviewees, however, mentioned that the Level 3 was not sufficient and that a Level 4 might indeed be better. In Samuel’s words:

Come on! Honestly, 3 is the passing mark, right? Everyone can get 3. It is just how many 4s and 5s you have got that matters (Samuel).

Terri, a new teacher who obtained an overall Level 4 at her second attempt at the LPATE, believed that English teachers in high-ability Band 1 secondary schools should at least reach Level 4. Trudy, a secondary English head of department, was of like mind, believing that English teachers should at least get Level 4 if they were to be seen as guaranteeing the quality of their teaching. The primary head of department Helena considered sufficient English language proficiency to be no less important in primary than in secondary schools, as primary schools were where students first encountered English, with it being crucial that such contact should constitute quality contact.

The counter argument was that the LPATE qualification was less important than was teachers’ ability to use English in their day-to-day teaching and encounters with speakers of English. Parents and children would be convinced by teachers’ language proficiency if teachers were able to talk to students and parents fluently and proficiently (Pamela).

The final point relating to English teachers obtaining Level 3 was that as teachers could now apply for exemption if they held a relevant degree in English, most English

teachers had not consequently sat the LPATE. Taking the LPATE would therefore appear to have greater relevance for aspiring English heads of department since these candidates required LPATE Level 4.

Level 4 for English Language Heads of Department

The first comment concerning heads of department attaining Level 4 was that heads of department needed to be seen as role models for the whole department. Drawing on her own experience of working with a number of heads of department across schools before the LPATE was introduced, Pamela commented on how she had encountered a number of heads of department whose language proficiency was inadequate; they had been appointed because of seniority. The LPATE requirement might at least ensure that heads of department had sufficient language proficiency to cope with the everyday work associated with running comparatively large departments of up to 15 English language teachers.

A head of department of a subject area, as Samuel stated, must be one of the most able in that subject area, if they are to be respected as leaders in a department. One of the heads' key duties was to give guidance to department members and to deal with the many documents that came in; without sufficiently high language proficiency, heads would struggle to fulfil their duties, and it was suggested (Trudy). Trudy elaborated how, in her school, all circulars sent out to parents were written in English, with it being the head of department's job to ensure that such documents were error free. Stella stated,

The head of department has to deal with a lot of documents and people. Sometimes when teachers have problems, you need to show them what to do. I think it just makes sense to me that you should do better than other teachers to be a leader, right? As an English head of department, I think you need to prove your proficiency in English. And that will be a very basic requirement, I guess (Stella).

The responses above show that English language heads of department should have both sufficient subject knowledge and good English language proficiency; otherwise, they may have a difficult time supporting the English language teachers in their panel. Without high proficiency, heads of department may not make sound judgments, for example about whether test questions are well set or certain textbooks are appropriate to student level (Tony).

While English language proficiency was considered to be important, other skills such as leadership, managerial, interpersonal communication and pedagogical skills were also cited as being important (e.g. Tony, Samuel, Helena, Stella and Scott). Tony noted that English language proficiency and the above-mentioned managerial-oriented skills were not mutually exclusive.

Participants who disagreed with English language heads of department having to obtain Level 4 considered a variety of other factors to be more important than merely English language proficiency and that achieving Level 4 and being a good head of department were not necessarily directly equivalent. Peter stated:

I'm not sure that if you can get an overall Level 4, it means that you have the ability to be an English Head of department. I mean that sometimes it's not quite equivalent, because the test is not made to test whether a person can be an English language head of department (Peter).

In this context, it may well be the case therefore that high levels of language proficiency and other skills are more co-dependent than mutually exclusive.

Peter recalled how in his school some very capable teachers refused to be head because they did not want to take the assessment. Sophie supported this view that in some schools, teachers were promoted to head of department simply on the basis of having obtained Level 4. Sandra considered that Level 3 should be sufficient, provided that it could be proved that this level had not been set too low and that she saw other qualities of managing the department as possibly being of greater or equal importance.

The examination nature of the LPATE made it difficult for some heads of department to pass; some participants observed. Selena, an English head of department in a Band 1 school, had herself failed the Speaking Test twice. Selena remarked that she felt confident about her oral English ability and she had no idea why she had failed the Speaking Test. Selena considered it was unfair to say the school had been wrong in promoting her to head, as she believed she had sufficient language proficiency and team management skills to be a successful English language head of department. The fact that not all English language heads of department held Level 4 was also complained about—on the grounds that the LPATE policy was not consistently enforced across schools (e.g. Scott)—although it should be remembered that holding a Level 4 qualification was not compulsory in terms of government policy.

Interestingly, in the current study, English heads of department holding negative views towards the LPATE tended to be those who had not taken the LPATE or who had not received what they perceived as a satisfactory result in the LPATE.

Discussion

Following a summary picture of interviewee responses, the discussion now examines certain key issues, as Section IV comes to a close.

Table 17.4 first provides a snapshot of the responses of stakeholders to the ten themes discussed in this chapter, revealing a summary of the views of principals, heads of departments and teachers.

In the table, Column 3—Number of participant responses—presents the total number of participants who commented on a particular theme. Since there are ten themes, some with subheadings, different fonts and bracketing has been used to aid readability. The major totals for the ten themes are in bold font. With the themes which had subheadings—e.g. themes (3) and (6)—subtheme totals are offset, centre right in round brackets. The final theme, (10), had sub-subthemes 'in favour', 'against' and 'neutral'. These sub-subthemes are offset to the right in square brackets.

Table 17.4 Summary of stakeholders' responses to the different themes

	Themes	No. of participant responses	Participants		
			<i>Ps</i> (N = 4)	<i>HoDs</i> (N = 13)	<i>Ts</i> (N = 7)
1	The LPATE—a response to a problem in the past	13	3	7	3
2	The LPATE and teaching as a profession	11	1	7	3
3	The LPATE and language proficiency	32			
	3.1 Importance of language proficiency	(11)	1	6	4
	3.2 The LPATE as a benchmark test	(21)	4	11	6
4	The language proficiency of Hong Kong English teachers	16	2	9	5
5	The LPATE and subject-matter knowledge	18	3	12	3
6	Pedagogical knowledge and skills	24			
	6.1 The classroom language assessment component	(9)	1	6	2
	6.2 Pedagogical knowledge and teaching skills	(15)	3	11	1
7	The negative side of the LPATE	8	2	5	1
8	Limitations of the LPATE	22	4	13	5
9	Changes over 14 years	13	3	8	2
10	LPATE requirements of English language teachers in Hong Kong	66			
	10.1 The need to take the LPATE	(18)	3	11	4
	10.2 The language proficiency Level requirements				
	10.2.1 Level 3 for English language teachers	(24)			
	In favour	[19]	3	10	6
	Against	[5]	1	3	1
	10.2.2 Level 4 for English language heads of department	(24)			
	In favour	[15]	2	7	6
	Against	[4]	0	3	1
	Neutral	[5]	2	3	0
Totals		223	38	132	53
			(17.0%)	(59.2%)	(23.8%)

Ps principal; *HoDs* heads of department; *Ts* teachers

Note Numbers in brackets indicate subtotals under a larger theme

As reported previously, the ratio of principals: heads of department: teachers was 4:13:7 (16.6%: 54.2%: 29.2%). In the light of this ratio, the spread of comments from participants in Table 17.4 (17.0%: 59.2%: 23.8%) can be seen to quite closely parallel participants' different backgrounds/positions.

When participants discussed the impact of the LPATE on the teaching profession, the two key aspects addressed were that the LPATE ensures English language standards, and the impact the LPATE had had upon subject-matter knowledge, and knowledge and awareness of grammar, in particular. Pedagogical skills—not a major focus in the LPATE—were also believed to be important to the teaching profession. The message that English language teaching is a profession that requires expertise in language skills, language knowledge and pedagogy (Richards, 2010) resonates strongly in the current study.

The foremost impact of the LPATE centred on improving language standards of English language teachers generally, by preventing those who were unqualified from entering the teaching profession. Teachers with high language proficiency would expose their students to quality language input (Andrews, 2003), would be more comfortable communicating with students in English and would have access to a wide vocabulary. Although the current study has not attempted to collect direct evidence of improvements in the language standard since it has been a study of perceptions, there is a strong belief that English language teachers should have high language standards and that minimum, agreed language standards have indeed been ensured by the assessment. The introduction of the LPATE was an attempt to address a problem Hong Kong faced before 2000 that English language teachers were not trained either content-wise or professionally in English (Tsui, Coniam, Sengupta, & Wu, 1994). This notwithstanding, however, the LPATE has come to be regarded as less relevant nowadays than it was in 1994 (Tsui et al., *ibid*) probably because more and more English teachers are exempted from the LPATE because they have relevant degrees and professional training.

The responses from key stakeholders—English teachers, English language heads of department and school principals—have demonstrated that in addition to purely improving language standards, a major contribution of the LPATE policy has been that of raising stakeholders' awareness that English teachers need to be subject-trained as well as proficient in subject-matter knowledge. The findings support the view that second language teachers are expected to have both 'general language proficiency' and 'academic proficiency'; i.e., they need to be proficient in reading, listening, writing, speaking and subject-specific knowledge (Elder & Kim, 2014).

While views differed on the issue of assessing grammar in the LPATE, the general perception was that—regardless of student level—English teachers need to be proficient in their knowledge of the English language and use English appropriately (Kamler, 1995). Such a view reinforces the importance of subject-matter knowledge of English language teachers (Borg, 2001; Myhill, Jones, & Watson, 2013) and echoes the impetus for the introduction of the LPATE—that English language teachers should be professionally and subject-trained.

Interviewees had different views towards assessing grammar in the LPATE assessment. Those who supported the assessment of grammar believed that regardless of the levels of students, English teachers need to be proficient in their knowledge of the English language and use English appropriately (Kamler, 1995). Such a view reinforces the importance of subject-matter knowledge of English language teachers (Borg, 2001; Myhill, et al., 2013) and echoes the impetus for the introduction of the LPATE that English language teachers should be both professionally and subject-trained. For those who saw knowledge of grammar as not being a key factor, the main argument was that the communicative approach paid more attention to expressing meanings and teachers could always consult grammar books if they had any questions of grammar. Although there is no sound evidence for the reasons for such a perception, it appears from the data that respondents with such a perception were less confident in their knowledge of grammar.

Overall, the picture that is painted by the respondents' perceptions of the LPATE and its impact is a positive one. They have quibbles with some of its aspects, but, on the whole, they approve of it setting and maintaining standards, requiring a Level 4 for heads of department and raising teachers' professional status.

Future Research

Having thoroughly researched stakeholders perceptions of the LPATE and its impact on teachers and society, there are a number of research areas that might be fruitfully explored.

The key factor here would involve an attempt to measure whether or not the introduction of the LPATE did raise English language standards among teachers of English. Such a study would be difficult but worthwhile, given that such a study would entail a large number of factors: the LPATE itself, the rise and growth of degree courses, the rise and growth of English teacher development courses (including the effect of immersion courses), the effect of exemptions on the teaching profession and the perceptions of other key stakeholders such as administrators, Education Bureau officials and a larger number of principals.

Conclusion

This chapter has complemented the findings of the quantitative study presented in Chap. 16. It has examined, via the software *NVivo*, the qualitative data gleaned by interview respondents, selected from those who had taken the quantitative survey. Their views were sought in a number of areas that arose from the initial quantitative data analysis in order to probe more deeply what these key stakeholders felt about the impact of the LPATE some years after its introduction.

Major findings were that the LPATE contributed to the raising of teacher language standards, enhanced a growing sense of professionalism among teachers, confirmed that heads of departments should have higher levels of LPATE scores but that, in a sense, the need for the LPATE has diminished in recent years as more and more teachers become subject and professionally qualified, thus allowing them to be exempted from the LPATE.

Notes

1. The research reported in Chapter 17 was supported by the Hong Kong Research Grants Council (grant number 18401514).
2. The change from two assessments to one was announced by the HKEAA in the September 2010 version of the Candidate Handbook. It was stated therein that the majority of CLA visits would be conducted by only one assessor although a commitment was retained to a small number of CLA assessments involving two visits. This change to the original structure of each CLA of a candidate being on the basis of two assessors making visits to two classes meant that, from 2010, 40% of candidates would get a second Classroom Language Assessment Classroom Language Assessment (CLA) by a different assessor.
3. From 2002 to 2014, under the aegis of the Common English Proficiency English proficiency Assessment Proficiency assessment Scheme (CEPAS), UGCUGC paid for graduating students in Hong Kong to take IELTS International English Language Testing System (IELTS) as a form of exit graduation 'indication' of language ability.

References

- Andrews, S. (2003). Teacher language awareness and the professional knowledge base of the L2 teacher. *Language Awareness, 12*(2), 81–95.
- Borg, S. (2001). Self-perception and practice in teaching grammar. *ELT Journal, 55*(1), 21–29.
- Elder, C., & Kim, S. (2014). Assessing teachers' language proficiency. In A. J. Kunnan (Ed.), *The companion to language assessment*. Wiley-Blackwell: Malden MA.
- Kamler, B. (1995). The grammar wars: Or what do teachers need to know about grammar. *English in Australia, 114*, 3–15.
- Myhill, D., Jones, S., & Watson, A. (2013). Grammar matters: How teachers' grammatical knowledge impacts on the teaching of writing. *Teaching and Teacher Education, 36*, 77–91.
- Richards, J. C. (2010). Competence and performance in language teaching. *RELC Journal, 41*(2), 101–122.
- Tsui, A. B. M., Coniam, D., Sengupta, S., & Wu, K. Y. (1994). Computer-mediated communication and teacher education: The case of TELENEX. In N. Bird, P. Falvey, A. B. M. Tsui, & A. McNeill (Eds.), *Language and Learning* (pp. 352–369). Hong Kong: Government Printer.

David Coniam is Chair Professor and Head of Department of the Department of Curriculum and Instruction in the Faculty of Education and Human Development at The Education University of Hong Kong, where he is a teacher educator, working with teachers in Hong Kong primary and secondary schools. His main publication and research interests are in language assessment, language teaching methodology and computer-assisted language learning.

Peter Falvey is a teacher educator, formerly a Head of Department in the Faculty of Education, The University of Hong Kong. His main publication and research interests are in language assessment, first and second language writing methodology and text linguistics.

Yangyu Xiao is a Senior Research Assistant in the Department of Curriculum and Instruction at The Education University of Hong Kong. Her publication and research interests are in the fields of formative assessment, language curriculum and teacher education.

Part V

Conclusion

Peter Falvey and David Coniam

The purpose of this part is twofold. Part 1 is comparatively short and recaps the various sections that make up the book in order to orientate the reader to the more investigative elements that constitute Part 2.

Part 2 consists of four parts and describes the constraints, weaknesses and strengths of the benchmark project, ending with a conclusion to the chapter.

Chapter 18

Concluding Comments

on the Benchmarking (LPATE) Project: Strengths, Weaknesses and Constraints



Peter Falvey and David Coniam

Abstract This chapter, in two parts, makes connections between and draws conclusions from the range of perspectives and lessons learnt from the exhaustive description and analysis of the LPATE initiative throughout the book. Part I begins with a brief summary of the ground covered in Sections I–IV. A useful timeline for this process from April 1996 to March 2000 (when the second administration of the LPATE took place) is provided by Urmston in Chap. 12, Fig. 12.1. Part II, which is grouped under the four main headings of Constraints, Weaknesses, Strengths and Conclusion, assesses the effectiveness of the LPATE within the context of educational reform and the specific socio-political context of Hong Kong in transition from British to Chinese control. The main findings, issues and lessons to be learned that arose throughout the 20 years of the LPATE are discussed.

Part I

A range of perspectives, from the developers of the original LPATE (Coniam and Falvey), to local university CHUK course providers (Mak and Xiao), the HKEAA (Urmston and Drave), the revised LPATE process (Urmston), the media (Drave) and a UGC-funded research project investigating the perspectives of other stakeholders (Coniam, Falvey and Xiao) have been described in this volume. They chart the development from the first initiation of the benchmark project in 1996 all the way through to a research project initiated almost twenty years after the initial decision to create language benchmarks for language teachers of English, Chinese and Putonghua in the Hong Kong school system. This chapter makes connections between and draws conclusions from the range of perspectives and lessons learnt from the exhaustive

P. Falvey (✉) · D. Coniam
Department of Curriculum and Instruction, Faculty of Education and Human Development,
The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, Hong Kong
e-mail: falvey@eduhk.hk

D. Coniam
e-mail: coniam@eduhk.hk

description and analysis of the LPATE initiative. It begins with a brief summary of the ground covered in Sections I–IV. A useful timeline for this process from April 1996 to March 2000 (when the second administration of the LPAT took place) is provided by Urmston in Chap. 12, Fig. 12.1.

The LPATE (Language Proficiency Assessment for Teachers of English), as the benchmark initiative was named in 2001, began with the determination of the Hong Kong Education Commission of ensuring:

1. That minimum language proficiency standards should be met by all teachers in their chosen medium of instruction.
2. That levels of language and professional competence ('benchmark' qualifications) should be established for all language teachers.

In early 1996, the Education and Manpower Bureau (EMB) tendered for the establishment of benchmarks for teachers of the English, Putonghua and Chinese languages for the following purposes:

- To establish benchmarks for primary teachers/secondary teachers/tertiary educators,
- To establish benchmarks for language teaching purposes/for promotional purposes,
- To establish benchmarks for teachers of subjects other than English language (i.e. teachers of such content subjects as physics, history, mathematics) who use English as the medium of instruction.

During the development of the LPATE, the government dropped the requirement for benchmarking tertiary educators and teachers of other subjects through the medium of English in order to prioritise the implementation of benchmarking for language teachers in schools.

Section I contextualised the development of LPATE and the academic literature on high-stakes assessment, benchmarking and teacher certification. The development of the LPATE—between 1996 and 2001—generated a large amount of data, assessment development and piloting events. These are chronicled and discussed with reference to the methodological approaches and analytical tools used to investigate them in Appendix A 'Methodological Approaches and Analytical Tools' in Chap. 8 at the end of Section I.

Section II described courses developed by The Chinese University of Hong Kong (CUHK) as an example of Hong Kong Government-funded developmental and certification programmes for teachers. These were developed and run at a number of local and overseas institutions, mainly in response to the early and hostile reaction to language benchmarking from the teachers' union.

Section III began by outlining the HKEEA's role in the development and administration of the LPATE from 2001 to 2006. It then described in detail revisions to the original LPATE between 2006 and 2008 when a new version was implemented. After outlining the process of maintaining standards for the three LPATE papers which use cut scores, the section closed with an examination of public reaction in the media to the development and implementation of the LPATE and the impact this had on Government decisions.

Section IV described stakeholders' perceptions of the impact of LPATE after 14 years.

Section V makes connections between and draws conclusions from the range of perspectives and lessons learnt from the exhaustive description and analysis of the LPATE project in previous sections.

Part II

Part II assesses the effectiveness of the LPATE within the context of educational reform and the specific sociopolitical context of Hong Kong in transition from British to Chinese control. The main findings, issues and lessons to be learned that arose throughout the 20 years of the LPATE are discussed. Part II is grouped under four main headings:

1. Constraints
2. Weaknesses
3. Strengths
4. Conclusion

Constraints

The compulsory introduction of a benchmarking assessment programme for a body of teachers—some of whom were already qualified, some with vast experience and extensive track records of successful employment, all of whom were already working, and where failure to meet the benchmark would lead to redundancy—was an ambitious, high-stakes, organisational change programme. In assessing the success or otherwise of the LPATE, it is first worth adding context from the literature of organisational change programmes and educational reform. As Williams (2014) points out:

Most change management programs initiated by leaders in organisations fail. They fail fundamentally because it is conceived as an outside-in process, moving about parts of the organisation, rather than an inside-out process which focuses on change within individuals. (P. 1)

Further, the difficulty of stating unequivocally whether reforms work or not are encapsulated in Terhart's (2013) citation of Blankertz (1977):

Decades ago, the German educational expert Blankertz (1977) concluded that from an educational point of view success and failure of education reform can hardly be assessed. It is not possible to make one decisive summative **final** evaluation. (P. 497)

Terhart (*ibid*) concludes by stating that, inevitably:

Once a programme or reform is under way, and especially if it is large and prestigious, it seems to become unstoppable even if problematic side effects soon become apparent. Too much energy, hope, resources or prestige have been invested and now it is simply too late. Even if it is senseless or harmful, the reform will not be stopped. Everyone has to continue – whatever the case. And so the reform gives birth to the necessity for another. It looks as if we will have to accept this as being normal. (P. 498)

Further, the original terms of reference did not include any form of impact measurement or request identification of criteria by which the success of LPATE could be assessed. The Hong Kong Government has resisted requests for release of full data for candidate numbers and associated pass rates on the different examinations and the university-run support courses. In fact, the closest indication achieved as to whether the LPATE project worked only came years later from the post hoc survey described in Section III of this volume.

So, organisational and educational change management programmes are characterised as difficult to measure, likely to fail and generative of further similar programmes. Most frequently cited reasons for failure relate to the management of stakeholder relations and include:

Trust: Government/teacher relations are rarely characterised by trust. The following quotation from a study of education in the UK by Mortimore and Mortimore (1998) could equally apply to Hong Kong in 2017.

In England education is high on the political agenda. The government's public pronouncements, however, have done little to lift teachers' morale at a time when requests for early retirement are rising and recruitment to the profession is falling... If the commonly held goal of raising standards is to be achieved, there is an urgent need to improve relations between those charged by the electorate to provide political leadership in education and those whose role it is to implement policies-and without whose support the most inspiring leadership will come to nought. (P. 1)

Dialogue: Teachers often feel that government interacts with them poorly, consults superficially or not at all. For example in England in 2013, the National Union of Teachers (NUT, 2013) protested strongly at not being consulted in response to the draft *National Curriculum* framework. Contrasting the government's approach with the consultation that the NUT had carried out with its own members, during which it received over 2000 responses, the NUT stated:

It is disappointing that the Government did not take the same approach [as us] and involve the profession directly in the formulation of its National Curriculum proposals (<http://www.toomuchtoosoon.org/national-curriculum-proposal-responses.html>).

Time: Governments frequently overestimate the pace at which change can occur, underestimating the time needed to carry out consultancy tasks (see Ben-Zur & Brezniz, 1981; Zakay, 1993) and the impact that the frequency and volume of change have on stakeholders.

Vision: Even where stakeholders are aligned over the goals, they will often differ as to the best means of achieving them; however, in the majority of cases there is a lack of alignment over goals for change, particularly in the educational arena. Advocates of change such as academic researchers, think tanks and politicians, who dominate the literature, often characterise resistance to change as resistance to innovation:

This is the problem of all theories and strategies of organisational change in institutions: the bosses want change, but those who will have to carry out the work lower down do not want change – and the change they want they do not get from their bosses. All theories of change management, organised change, organisational development and so on ultimately circle around this theme of resistance to innovation. (Harvey & Broyles, 2010)

Against this backdrop of factors that usually militate against success for educational reform, the peculiar circumstances pertaining to Hong Kong in the process of transition from a colonial British government to Chinese rule as part of the Special Administrative Region must be added. The Professional Teacher Union (PTU) was set up in 1974, in response to a 15% cut in salaries of certified teachers. Relations with the British colonial government were never particularly good and as the PTU grew in influence—both as a trade union and as a political force—the government viewed it as a thorn in its side [Note 1]. While the PTU was sceptical of the colonial government’s commitment to take steps to protect professionalism, democracy and justice in the lead up to the handover, it likewise had no reason to believe that they would be respected post-handover by the incoming Chinese administration.

Within this context of poor relations, in 1996, the Hong Kong Government, while only one year away from the handover of Hong Kong to China, chose to initiate a raft of educational reforms including the benchmarking project described in this volume. Providing examples of these educational reforms, Drave, in Chap. 16, says that they included: school-based assessment; quality assurance; the launch of new degree programmes in teacher training institutions and important and controversial decisions about the medium of instruction [Note 2]. Post-handover, Cheng (2009) reports that the HKSAR Government increased the pace of change by introducing

an important blueprint for the educational development of Hong Kong in the new century which included a thorough review of the structure of pre-primary, primary, secondary, and tertiary education, as well as the school curriculum and examination system, while the Board of Education had at the same time completed a review of 9-year compulsory education (Board of Education, 1997). (Cheng, 2009:67)

Cheng (*ibid*) summarises these reforms below:

- Reforming the admission systems and public examinations so as to break down barriers and create room for all,
- Reforming the curricula and improving teaching methods,
- Improving the assessment mechanism to supplement learning and teaching,
- Providing more diverse opportunities for lifelong learning at senior secondary level and beyond
- Formulating an effective resources strategy,
- Enhancing the professionalism of teachers,
- Implementing measures to support frontline educators.

Cheng (*ibid*) cites reports from the Hong Kong Mood Disorders Centre, Hong Kong Federation of Education Workers and Hong Kong Professional Teachers’ Union showing that that the major sources of spiritual pressure and work pressure on teachers are derived from the changes in this period, namely implementation of

educational reforms (88–97%), school administrative work (65–96%) and additional requirements of professional training (62–90%).

In addition to the pressures exerted on teachers by this raft of qualitative changes, there were two demographic factors that threatened their job security, particularly among primary school teachers. A falling birth rate and the cessation of wholesale illegal immigration from China was resulting in a decline in student numbers and school closures. Cheng (*ibid*: 82) points out that the accumulative pressure on teachers led to protest and ultimately recognition by government:

After the protest of over 10,000 teachers at the beginning of 2006, the Government began to understand the serious negative impacts of educational reforms on teachers **and schools** and immediately announced nine measures in a total of 1.8 billion HK dollars to address the issues of high work pressure on teachers. Also a committee on studying the work pressure on teachers was established to investigate the details of problems and recommend the solutions. (P. 82)

This is the sociopolitical context surrounding the implementation of the benchmarking programme. As Drave's analysis in Chap. 17 showed there was clearly confusion around the range of reforms leading to many of the comments and criticisms of the LPATE in the press, often written by teachers who appeared to be ill-informed. Teacher reaction to the LPATE was strong because as well as being worried about their English language competency, and the thought of having to undergo what they perceived as both stringent testing and a threat to their jobs, they also felt their integrity was being attacked. The survey on the impact of the LPATE, described in Section IV, revealed that, 20 years on, the feelings of distrust persist. The teachers felt humiliated because they had had to take a test in their mid-fifties if they wished to keep their jobs (Victoria). This feeling of humiliation was closely linked to not feeling trusted. Hope commented thus:

I would use the word 'humiliating'. I find it humiliating. That means you don't trust me as an English major. My English should be OK. At least I could communicate with my students and could discuss educational matters with my colleague, so I really felt very bad about the idea of being asked to take the LPATE at that time, because I think A-Level would be quite an accurate estimation.

One of the respondents cited in Chap. 17, Hugo, a head of department, encouraged all teachers in his department to take the LPATE in order to showcase their language standard. However, Honey and Harriet were strongly opposed to such a move, stating that requiring teachers to take the LPATE showed how heads of department did not trust their teachers.

As a result, teachers urged their union leaders to protest against the benchmark initiative. There were a number of demonstrations and other protests carried out to try to put pressure both on the colonial government and the incoming government. As Urmston notes in Chap. 12, pressure exerted by the PTU eventually led to compromises by the Government, e.g., the introduction of teacher development courses and the establishment of exemptions from the LPATE, as well as subsequent changes to parts of the assessment that were perceived to be difficult.

Weaknesses

In considering what could have been done differently, one factor stands out. With hindsight it is clear that rather than trying to impose LPATE on teachers, the British colonial government and the Chinese authorities that succeeded them could have avoided some of the protests, conflict and the subsequent watering down of the scope and rigour of the benchmark programme had they succeeded in getting the PTU and teacher opinion on board as early as possible in the design process.

In 1966, the consultants carried out a territory-wide survey of the views and attitudes of teachers of English towards language benchmarks for teachers of English. The results of the survey showed overwhelming support for benchmarking standards. In addition to both the colonial government and the incoming administration, the influential Hong Kong business community supported professionalising English teaching—seeing it as the key to improving English generally and therefore key to business and commercial success. As such, the goals of the LPATE, namely professionalisation and the raising of English language standards among teachers were shared by all key stakeholders. There was the chance to promote a policy which, if handled well, could have gone some way towards building trust between government and teachers, at least with the incoming regime. In addition to clearer communication and the creation of a collaborative approach to reform, a risk analysis during the late 1990s of the number and profile of teachers who may have been in danger of failing to meet the benchmark should have been carried out as this could have allayed fears. It would also have made sense to pre-empt the exemptions that would later be conceded, together with the initiative to provide funded wrap-around training and certification although, as has been described in Section I, the consultants felt that sufficient development progress had not been made to outline what sorts of exemptions could be allowed.

Achieving the above would necessarily have required more time, and the outgoing colonial British government was perhaps naïve in its setting of deadlines for LPATE and other reforms. As a result of the tight timeframes, the original scope of the benchmarking programme was reduced. The ACTEQ originally stipulated the following should occur:

- Establishing benchmarks for primary teachers/secondary teachers/tertiary educators,
- Establishing benchmarks for language teaching purposes/for promotional purposes,
- Establishing benchmarks for teachers of subjects other than English language (i.e. teachers of such content subjects as physics, history, mathematics) who use English as the medium of instruction.

However, concerning the first bullet point, benchmarks were established for only the first two categories, primary and secondary teachers, not for tertiary educators. Regarding the third bullet point on benchmarks for teachers of subjects other than English language (i.e. teachers of such content subjects as physics, history and mathematics), the initiative was dropped. The failure to create these language benchmarks is

most unfortunate and can be considered a missed opportunity as the use of Cantonese in content subjects was prevalent in many schools in many subjects and contributed to the medium of instruction decisions described in Note 2.

Finally, in terms of programme design, it would also have been helpful to include a full impact survey of the project. Although interviews were carried out with teachers by the HKEAA, these were limited in number. As mentioned previously, the closest the authors got to discovering whether the project really worked only came years later from the post hoc survey described in Section IV.

Moving the discussion to during and after the 2006–2007 revision, a number of discussion points arise.

First, the timing of the LPATE revision process appears curious, given that all serving teachers of English would have been benchmarked by LPATE, development courses, or exemption by the end of 2005. The LPATE would no longer be administered to teacher education candidates in tertiary institutions although it has been administered subsequently to (self-claimed) serving teachers. It would be reserved for those who came from overseas without both academic and professional qualifications and others who did not fall into the category of serving teachers. Turning to the revisions themselves, it is somewhat unfortunate that the reading aloud of a poem was eliminated because of the complaints of some who said it was too difficult for teachers to master or too difficult to find appropriate. It should be accepted that reading aloud, particularly the reading aloud of poetry, is a key skill that native speakers as well as other speakers of English must master if they are to teach language arts effectively as part of their English classes. The fact that it is difficult to master is even more of a reason to retain it rather than cut it from the assessment battery. As Urmston states in Chap. 13, when pointing out that the assessment of reading a poem was a good discriminator of ability:

The choice of a poem as a text in addition to a prose passage was controversial **but was found to be a good discriminator between candidates of differing abilities and provided assessors with the opportunity to measure each candidate's ability to deal with different types of text**. Assessors were made aware of the fact that reading poetry aloud is a difficult skill, even for native speakers, and this was borne in mind during the assessor standardisation process and during the assessing itself.

In addition, the fact that the exam setters found it a bit difficult to find appropriate poems should not have been accepted. There are millions of poems in the English language—all that is required is a certain amount of work to find adequate ones. Wikipedia contains a comprehensive list of poets who have published in English (https://en.wikipedia.org/wiki/List_of_English-language_poets, accessed November 2017) to say nothing of those whose poetry has been translated into English.

Evidence for the efficacy of reading aloud can be found in Anderson et al. (1985, p. 23) who state unequivocally that reading to children is ‘the single most important activity for building the knowledge required for eventual success (in learning to read)’. Evidence can also be found in Lane and Wright (2007), Chalfant (2013), Hoffman et al. (1993), and Rasinski (2017). It should be noted that the sources cited above all focus on native speaking students. However, Amer (1997), speaking in an ELT context states:

Although reading aloud receives considerable emphasis in English as a first language, it is traditionally discouraged by EFL teachers and methodology specialists. Reading aloud, in fact, is particularly important for EFL learners at the early stage of learning. Beginning readers tend to read word by word. Reading aloud helps them read larger semantic units rather than focusing on graphic cues. (p. 43)

In addition, work published in Hong Kong towards the end of the twentieth century stressed the importance of reading aloud, including the reading aloud of poems (c.f., Falvey, 1997; Harris & Leung, 1997; Tyrell, 1997). These authors all focused on non-native speaking language students.

In the context of the assessment of writing, it would also appear unfortunate that some of the scales were eliminated. On consideration, it might have been better to retain criterion-referenced assessment for all the Writing Test components. A performance test demands a criterion-referenced approach. The fact that some informants to the revision committee objected that parts of the Writing Test were difficult is not a valid reason for eliminating some of the scales and descriptors which provide much better feedback for test-takers than numerical scores on a test. Within a now outdated paradigm, numerical scores might be considered reliable but scales and descriptors would certainly be considered more valid. Objections to some parts of the writing assessment instrument may have arisen because the notion of criterion-referenced assessment was not, at the time, well known by classroom teachers, nor did they realise that useful and beneficial feedback could be provided rather than a raw score for the elements that supplanted the criterion-referenced section.

In the context of the Listening Test revision, it was reported in Chap. 6 that the proposed use of video material for the Listening test was abandoned because of technical difficulties. Since the first trial of the first LPATE pilot listening tests approximately 20 years ago, studies on multimodality have revealed how information presented in multiple modes may impact on the comprehension of information. Audiovisual materials have been promoted and are considered to enrich language learning (Vanderplank, 2009). A study comparing second language university learners' comprehension of an authentic BBC audiovisual recording reveals that comprehension improves when learners are exposed to a text in several modalities (Guichon & McLornan, 2008).

Whereas it can be concluded that visual input is likely to support comprehension, some research studies show that video input itself does not have obvious advantages. Feak and Salehzadeh (2001) implemented video language assessment with non-native English speakers. Students reported that they did not know whether or not they should watch the video and that they could not concentrate on listening. Coniam (2001), similarly, compared video and audio assessment for the Hong Kong English Language Benchmark Test at its initial trialling. There were no significant differences between scores of teachers taking audio assessment and teachers taking video assessment. The video-taking group did not feel they had gained any advantages and reported that they felt that they would do better without being distracted by images. It was for problems such as those cited above that the revision of the LPATE Listening Test did not pursue the hoped-for inclusion of video into the revised test.

Moving onto Classroom Language Assessment, it is to be regretted that the two classroom visits were changed to one only, no matter what methods were used to spot check the assessment by the use of some second visits. This can be considered a retrograde development and it is not entirely clear why it was carried out and approved. In Section IV, teachers have been quoted wondering about the validity of cutting back classroom observations from two to one. Since 2010, the Classroom Language Assessment component has been assessed only by one examiner in one single lesson. Helena (quoted in Section IV), quite rightly doubted whether one classroom observation was sufficient for a sound judgement of a teacher's language proficiency to be made.

There is little research evidence regarding how many lessons need to be assessed in order to make a valid assessment of a teacher's proficiency. Below well-known sources are cited and then information is provided about classroom assessment for the Cambridge CELTA and DELTA worldwide qualifications.

Harris Schools Solutions, a large company who act as consultants for counties and states in the USA state quote the principal of a large secondary school:

It has been my experience that there are never too many classroom observations. Actually, my teachers enjoy frequent classroom visits. Our teachers have expressed to us that they like for us to visit their classrooms. They have also told us that they like for us to stay for extended periods of time when we do visit. Harris School Solutions – (<https://harrisschoolsolutions.com/blogposts/how-many-classroom-observations-are-too-many/>, accessed November 2017)

Stressing the importance of a number of observations and assessments, Danielson, an American consultant on teacher education is unequivocal when commenting on classroom observations and their frequency:

One of the important findings of the MET study was that **the reliability of observations increased with both the number of observations and the number of observers**. While probably unrealistic for practicing Educators, four observations, conducted by several different observers, was about twice as reliable as an observation of a single lesson.

Overall, my recommendation is that **the observation component of a full evaluation (should) consist of one full lesson, and three additional, shorter observations, and that these observations are conducted by two different individuals**. (<https://www.danielsongroup.org/questions-about-observations-of-classroom-practice/>, accessed November 2017)

In terms of other assessment bodies, the British National Union of Teachers, after negotiation with government issued a classroom observation protocol. Seeking to defend their members' rights, they stated:

The Regulations place a maximum of three hours on classroom observation except where evidence emerges that gives rise to concern about a teacher's performance. (<http://www.teachers.org.uk/files/active/0/Observation4798.pdf>, accessed November 2017)

CELTA course regulations state that 'You will teach for a total of 6 h, working with adult classes at a minimum of two levels of ability. Assessment is based on your overall performance'. [Note 3]

In this instance, it should be noted that classroom assessment will be at a minimum of two levels of ability, e.g. Secondary 1 and Secondary 4 (Years 7 and 10). Also,

six hours would equate to six observations if the lessons lasted for one hour. In Hong Kong, however, lessons are shorter than one hour so there would be nine observations in six hours, even allowing for 40-minute lessons.

DELTA programmes run by Cambridge English Language Assessment are for teachers with at least one year's experience. In the classroom assessment module, there are five classroom observations carried out by three different assessors.

The assignments incorporate both background essays and observed teaching. The first formal observation is completed during the Orientation Course. The second, third and fourth observation are done by your Local Tutor. Your final assignment will be assessed by an external assessor. Assessors are experienced *DELTA* teacher trainers nominated by Cambridge. (<http://www.cambridgeenglish.org/teaching-english/teaching-qualifications/delta/>, accessed November 2017)

In comparison to the recommendations and standards cited above, the allocation of one assessed lesson only seems paltry and, apart from accusations of lack of reliability, can be faulted on validity grounds. Hence the rather strong views that the CLA component of the LPATE should consist of at least two observed and assessed lessons.

The perception of English language teachers using Chinese to teach English was reported in Section I as being one of the issues that led to the LPATE initiative. More recent research on topics such as 'using multilingualism in language teaching' and 'translanguaging' (see, e.g. Creese & Blackledge, 2010) have destabilised the earlier principle of 'monolingual teaching in the target language.'

Although teaching English through English or hiring native speakers to teach English has become increasingly popular, the proponents of a 'multilingual model of English Language teaching' consider that non-native English teachers or multilingual English teachers provide a more appropriate linguistic model than the native speakers (Kirkpatrick, 2010). With specific reference to using multilingualism in English-language classroom, Cummins (2009) strongly believes that bilingual instruction has more obvious advantages than monolingual instruction in that bilingual instruction works better at activating students existing knowledge which is encoded in their L1, strengthening students' translation skills which would help enhance their linguistic awareness, scaffolding their language output and enabling them to use high-order thinking skills. As Cook (2001) exemplified, there are several ways that L1 can be used positively and productively in a language classroom. Examples included teachers using L1 for checking meaning, for explaining grammar, for organising tasks, for maintaining discipline in the classroom, and for contacting individual students. Cook (2001) suggests that teachers should consider four factors if they would like to use L1 in the classroom: *efficiency* (can something be done more effectively through L1?), *learning* (will L2 learning be helped by using L1?), *naturalness* (do participants feel more comfortable about communicating a certain topic in their L1 rather than L2?) and *external relevance* (will the use of both languages help students master the L2 they need beyond the classroom context?).

While the studies and evidence cited above support, in principle, some use of the mother tongue in the classroom, research generally supports the use of the second language from the perspective of pedagogical effectiveness. In recent years, there

has been a growth in the popularity and uptake of ‘plurilingual’ pedagogies (Lin, 2013) as well as what has been termed a ‘translanguaging’ approach to the use of language(s) in the classroom (García & Wei, 2014). [Note 4]

Plurilingual pedagogies and translanguaging are encouraged in that such approaches are viewed as a move towards greater flexibility in using classroom languages, in light of the plurilingual nature of classroom interaction and communication repertoires (Lin, 2013). Lin (ibid) argues that—unlike the grammar-translation approach—a plurilingual approach focuses more on how teachers use local resources in scaffolding learning. García and Wei (2014) argue for the use of the L1 from the perspective of supporting student learning, in particular given the fact that in the classroom both teachers and students draw on complex resources for real-time meaning-making. García and Wei (Ibid)—strong proponents of translanguaging—make the case regarding the use in education of both the ‘home’ language (students’ L1) and the target language (students’ L2). García and Wei (Ibid) further argue against the practice of minimising (i.e. ignoring or avoiding) students’ home language and against the practice of keeping languages separate. Their arguments are, however, predicated on the issue that the L1 is used to support learning, rather than—as was the case in Hong Kong before the benchmark assessment was implemented—the L1 being used because of teacher language proficiency problems. Thus, from a plurilingual perspective, requiring language teachers to achieve L2 language proficiency at a benchmarked level is not in conflict with whether more than one language should be used in the classroom.

Successes

In this part, the successes of the LPATE are discussed. These broadly fall into two categories: first, achieving the result that was intended by the reform and associated benefits, and secondly contributions that the LPATE can make to benchmarking projects in other countries through the design and development process and the lessons learnt that are documented in this volume. The successes of the LPATE should be viewed in the light of the many constraints described earlier in this chapter, such as the generally poor success rate worldwide of educational reform and change management and the particular sociopolitical context of Hong Kong.

The LPATE was designed as a benchmark against which English teachers working in Hong Kong could be assessed, one which would discriminate between teachers who were required to meet a minimum standard of language competency (Level 3) and those who failed to do so. The test achieved this and by 2006 all working teachers in Hong Kong who were not granted exemption had been assessed. The test met stringent criteria for validity, rigour, integrity and reliability that would be expected for a high-stakes assessment and has also stood the test of time: it is still administered to potential teachers in Hong Kong who are not entering the profession following qualification at local teacher education institutions. Further, although some teacher education institutions in Hong Kong have adopted LPATE-

type benchmarking criteria and use them as part of their continuous and summative assessment of their students, it should be noted that not all do so; the Education University of Hong Kong (formerly the HKIEd) requires a score of 7.0 on the IELTS before undergraduate students of English language education can graduate. Level 3 on the LPATE criteria has clearly been accepted as a minimum requirement for teachers of English in Hong Kong. This position is supported by the impact research described in Chap. 18 where, prominent in interview respondents' comments, was the belief 'that the LPATE ensures English Language standards'.

Although not goals of the LPATE, two washback effects were the development of a sense of professionalism among teachers and a positive developmental impact on subject knowledge and classroom practice of Hong Kong teachers of English. Richards (2010) lists ten core dimensions of skill and expertise in language teaching:

1. Language proficiency
2. Content knowledge
3. Teaching skills
4. Contextual knowledge
5. Language teacher identity
6. Learner-focused teaching
7. Specialised cognitive skills
8. Theorising from practice
9. Joining a community of practice
10. Professionalism.

Of Richards' ten points, the LPATE impacted on the following eight for English language teachers in Hong Kong: language proficiency; content knowledge; teaching skills (in preparing for the classroom language assessment); contextual knowledge (how to present meaningfully to students—these were especially stressed in the teacher development programmes); learner-focused teaching (in preparing for classroom language assessment); specialised cognitive skills (from teacher development programmes); joining a community of practice (especially through the development courses); professionalisation.

The last point on the list—professionalism—is understood by Richards to be the accumulation of the previous nine points. In the context of Hong Kong and the LPATE, there is further evidence of professionalism or professionalisation of the sector in the adoption of Level 4 of the benchmarking criteria as a requirement for promotion to panel chair positions. Stella, an English language head of department stated:

The head of department has to deal with a lot of documents and people. Sometimes when teachers have problems, you need to show them what to do. I think it just makes sense to me that you should do better than other teachers to be a leader, right? As an English head of department, I think you need to prove your proficiency in English. And that will be a very basic requirement, I guess.

Turning to the skills of writing, one interesting finding in the teacher development courses run by The Chinese University of Hong Kong (CUHK) was that teachers

on the development courses did better in their writing assessments than did those who simply took the formal LPATE assessment which had a quite large failure rate. Notoriously, writing is the most difficult skill to acquire mastery over and the implication of the Chinese University findings is that slow, steady development, coupled with reflection and peer and instructor perseverance paid off. Further, if indicative of wrap-around courses offered by all institutions, the very positive feedback received by CUHK demonstrates that passing participants, as well as achieving minimum language standards, also improved subject-matter knowledge, knowledge and awareness of grammar and pedagogical skills. This is further corroborated in respondent feedback cited in Chap. 17:

When the interview participants discussed the impact of the LPATE on the teaching profession, the two prominent aspects addressed were: that the LPATE ensures English language standards, and the LPATE improved language subject-matter knowledge, and knowledge and awareness of grammar, in particular. The pedagogical skills which were not a major focus in the LPATE test-were also believed to be important to the teaching profession. The message that English language teaching is a profession that requires expertise in language skills, language knowledge and pedagogy (Richards, 2010) is strongly delivered in the current study.

The second area of potential success, attributable to the LPATE, is the possible contribution that the LPATE can make to benchmarking projects in other countries by them appropriating the design and development process and being aware of the lessons learnt, documented in this volume. The most significant lessons learnt cohere around stakeholder and change management. A first step in managing teachers as stakeholders is to know what their beliefs and attitudes are towards a proposed reform. In the case of the LPATE, this was achieved through a survey that had a response rate of above 90% among the approximately 12,500 teachers of English in Hong Kong and could provide a template for surveys of different teacher populations. Other lessons learnt include better communications with teachers and with the local media that take account of potential teacher resentment and fears and take steps to address them by allowing exemptions and guaranteeing funded wraparound development courses to assist teachers in achieving benchmarks.

The successful wraparound courses are also available as a template for other countries wishing to deliver a benchmarking programme and the feedback suggests that such courses allow for positive washback from the assessment programme to occur in terms of improvements, both in and out of the classroom, of the type listed by Richards (2010) above. The success of these courses can, to some degree, be attributed to the quality of the criteria that they were based on and these criteria are available for other countries to use or adapt (see Kimura, Nakata, Ikeno, Naganuma, & Andrews, 2017 for a description of a study using LPATE scales and descriptors with Japanese teachers).

In many countries, as was the case in Hong Kong, many people, especially non-educators, believe that the model English teachers should aspire to is that of the 'native speaker'. Teachers cited in Section IV stated that English teachers were generally regarded as being capable and effective classroom teachers but they were considered to be less proficient at using English in settings outside the classroom, such

as commenting on current affairs (Suzanna), expressing views in English in panel meetings (Sophie), and communicating with native speakers (Suzanna, Tania). These pronunciation comments brought out how RP or Standard American pronunciation was considered to be a more acceptable 'model' for Hong Kong English teachers. However, the model conceived as underpinning the LPATE was that of the 'educated Hong Kong speaker'. The reasoning behind the decision to use the 'educated Hong Kong speaker' was that:

1. It would be far too expensive to send all teachers overseas for an extended period to acquire a standard US accent or a Received Pronunciation (RP) accent.
2. Even if huge resources were allocated for the above purpose there would still be costly failures to achieve such an accent.
3. To become an educated Hong Kong speaker is attainable, internationally comprehensible and much less expensive to achieve.
4. As Davies (1995) stated: The native speaker is a fine myth: we need it as a model, a goal, almost an inspiration. But it is useless as a measure (p. 157).

These points were put to ACTEQ at an early stage and approved. They then were put to the relevant subject committees for the Speaking Test and Classroom Language Assessment components of the LPATE for approval. In many countries where a benchmarking programme of this kind may be implemented, similar misguided positive public attitudes towards a native speaker model may need to be addressed and the LPATE can be cited as evidence of the successful implementation of a local educated speaker model.

Conclusion

This chapter began with a brief summary of the volume before moving on to assess the impact of the LPATE. Beginning by outlining the constraints within which the LPATE programme was undertaken, the chapter went on to point out weaknesses of the LPATE and lessons learnt from its implementation. The final section looked at successes describing them in terms of how the LPATE achieved what it set out to do, produced other unintended benefits and, with the production of this volume which documents its impact over 20 years, shows how the LPATE could be a potential model for the implementation of benchmarking programmes in other countries.

Notes

1. As a trade union, the PTU was, by the late 1980s, the largest single union in Hong Kong with over 32,000 members, (Butenhoff, 1999). As a political force, the PTU has held the Educational functional constituency in the Legislative Council

- since its creation in 1985, continually seeking ‘Professionalism, Democracy and Justice!’ (<https://www.hkptu.org/english>).
2. The medium of instruction, although labelled English-medium, was, for many years a mixture of Cantonese and English. The Education Commission wished to do away with the hypocritical labelling of schools and stated that all secondary schools would teach through the medium of Chinese unless they could prove that they had the staff and resources to teach effectively through English. This decision provoked fury among many school principals who did not want the cachet of English-medium tuition to be taken away from them.
 3. CELTA courses are run by approved centres, based on specifications produced by Cambridge English (2017). All courses have a minimum of 120 contact hours.
 4. *Plurilingualism*—by which all are entitled to develop a degree of communicative ability in a number of languages over their lifetime in accordance with their needs—is being promoted by Council of Europe language education policies (see https://www.coe.int/t/dg4/linguistic/Division_EN.asp accessed 20 April 2018). *Translanguaging* is the act performed by bilinguals of accessing different linguistic features or various modes of what are described as autonomous languages, in order to maximise communicative potential (García, 2009, p. 140).

References

- Amer, A. A. (1997). The effect of the teacher’s reading aloud on the reading comprehension of EFL students. *English Language Teaching Journal*, 51(1), 43–47.
- Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, I. A. G. (1985). *Becoming a nation of readers: The report of the commission on reading*. Washington, DC: US Department of Education.
- Baum, H. S. (2002). Why school systems resist reform: A psychoanalytic perspective. *Human Relations*, 55(2), 173–198. <https://doi.org/10.1177/0018726702055002182>.
- Ben-Zur, H. B., & Brezniz, S. J. (1981). The effect of time pressure on risky choice behavior. *Acta Psychologica*, 47(2), 89–104.
- Blanchard, K. (2010). Mastering the art of change. *Training Journal*, 44–47.
- Blankertz, H. (1977). Was heißt: ein Bildungswesen ‘pädagogisch’ zu verbessern? [What is meant by: improving an educational system in ‘pedagogical’ respect?]. In J. Derbolav (Ed.), *Grundlagen und Probleme der Bildungspolitik [foundations and problems of educational politics]* (pp. 79–87). München: Pieper.
- Butenhoff, L. (1999). *Social movements and political reform in Hong Kong*. Westport, CT: Greenwood Publishing Group.
- Cambridge English. (2017). CELTA. <http://www.cambridgeenglish.org/teaching-english/teaching-qualifications/celta/about-the-celta-course/>. Accessed November 2017.
- Chalfant, P. (2013). *Teacher professional development and storybook reading*. Unpublished PhD Thesis. Florida: University of Florida.
- Cheng, Y. C. (2009). Hong Kong educational reforms in the last decade: Reform syndrome and new developments. *International Journal of Educational Management*, 23(1), 65–86.
- Coniam, D. (2001). The use of audio or video comprehension as an assessment instrument in the certification of English language teachers: A case study. *System*, 29(1), 1–14.
- Cook, V. (2001). Using the first language in the classroom. *Canadian Modern Language Review*, 57(3), 402–423.
- Creese, A., & Blackledge, A. (2010). Translanguaging in the bilingual classroom: A pedagogy for learning and teaching? *Modern Language Journal*, 94(1), 103–115.

- Cummins, J. I. M. (2009). Multilingualism in the English language classroom: Pedagogical considerations. *TESOL Quarterly*, 43(2), 317–321. <https://doi.org/10.1002/j.1545-7249.2009.tb00711.x>.
- Danielson, C. (2017). Questions about observations of classroom practice. <https://www.danielsongroup.org/questions-about-observations-of-classroom-practice/>. Accessed November 2017.
- Davies, A. (1995). Proficiency of the native speaker: What are we trying to achieve in ELT? In G. Cook & G. Seidlhofer (Eds.), *Principle and practice in applied linguistics*. Oxford: Oxford University Press.
- Falvey, M. (1997). Verse and worse: Poetry and rhyme in the EFL primary school classroom. In P. Falvey, & P. Kennedy (Eds.), *Learning language through literature: A sourcebook for teachers of English in Hong Kong*. Hong Kong University Press: Hong Kong.
- Falvey, P., & Kennedy, P. (Eds.). (1997). *Learning language through literature: A sourcebook for teachers of English in Hong Kong*. Hong Kong: Hong Kong University Press.
- Feak, C. B., & Salehzadeh, J. (2001). Challenges and issues in developing an EAP video listening placement assessment: A view from one program. *English for Specific Purposes*, 20(Supplement 1), 477–493. [https://doi.org/10.1016/S0889-4906\(01\)00021-7](https://doi.org/10.1016/S0889-4906(01)00021-7).
- Fountas, I., & Pinnell, G. S. (1996). *Guided reading* (2nd ed.). Heinemann: Portsmouth, NH.
- García, O. (2009). Education, multilingualism and translanguaging in the 21st century. In A. Mohanty, M. Panda, R. Phillipson, & T. Skutnabb-Kangas (Eds.), *Multilingual education for social justice: Globalising the local* (pp. 128–145). New Delhi: Orient Blackswan.
- García, O., & Wei, L. (2014). *Translanguaging: Language, bilingualism and education*. London: Palgrave Macmillan.
- García, O., Johnson, S., & Seltzer, K. (2017). *The translanguaging classroom. Leveraging student bilingualism for learning*. Caslon: Philadelphia.
- Guichon, N., & McLornan, S. (2008). The effects of multimodality on L2 learners: Implications for call resource design. *System*, 36(1), 85–93. <https://doi.org/10.1016/j.system.2007.11.005>.
- Harris Schools Solutions. (2017). How many classroom observations are too many? <https://harrischoolsolutions.com/blogposts/how-many-classroom-observations-are-too-many/>. Accessed November 2017.
- Harris, J., & Leung, M. (1997). The use of children's literature in the English primary classroom. In P. Falvey, & P. Kennedy (Eds.), *Learning language through literature: A sourcebook for teachers of English in Hong Kong*. Hong Kong University Press: Hong Kong.
- Harvey, T. R., & Broyles, E. A. (2010). *Resistance to change: A guide to harnessing its positive power*. New York: Rowman & Littlefield.
- Hoffmann, J., Battle, J., & Roser, N. (1993). Reading aloud in classrooms: From the modal towards a 'model'. *The Reading Teacher*, 46(6), 496–503.
- Hughes, J. (2007). Poetry: A powerful medium for literacy and technology development. In *What works?* Ontario: Literacy and Numeracy Secretariat: Ontario.
- IBM Corporation. (2008). Making change work. In L. Allan (Ed.), *Business Performance Pty Ltd*. <http://www.businessperform.com/change-management/change-management-faillure.html>. Accessed November 2017.
- Keller, S., & Aiken, C. (2008). *The inconvenient truth about change*. <http://projektmanazer.cz/kurz/soubory/modul-c/the-inconvenient-truth-about-change-management.pdf>. Accessed November 2017.
- Kimura, Y., Nakata, Y., Ikeno, O., Naganuma, N., & Andrews, S. (2017). *Language Testing in Asia*, 7(3). <https://doi.org/10.1186/s40468-017-0035-2>.
- Kirkpatrick, A. (2010). English as an Asian lingua franca and the multilingual model of ELT. *Language Teaching*, 44(2), 212–224. <https://doi.org/10.1017/S0261444810000145>.
- Lane, H., & Wright T. L. (2007). Maximizing the effectiveness of reading aloud. *The Reading Teacher*, 60(7), 668–675.
- Lin, A. (2013). Toward paradigmatic change in TESOL methodologies: Building plurilingual pedagogies from the ground up. *TESOL Quarterly*, 47(3), 521–545.

- Mortimore, P., & Mortimore, J. (1998). The political and the professional in education: An unnecessary conflict? *Journal of Education for Teaching: International Research and Pedagogy*, 24(3), 205–219.
- National Union of Teachers. (2017). *A classroom observation protocol: Guidelines for nut school representatives*. <http://www.teachers.org.uk/files/active/0/Observation4798.pdf>. Accessed November 2017.
- National Union of Teachers. (NUT). (2013). *Too much too soon*. <http://www.toomuchtoosoon.org/national-curriculum-proposal-responses.html>). Accessed November 2017.
- Neuman, S., Copple, S. B., & Bredekamp, S. (2000). *Developmentally appropriate practices for young children*. Washington, DC: National Association for the Education of Young Children.
- Palmer, J. (2017). Change management in practice: Why does change fail? <https://www.projectsart.co.uk/change-management-in-practice.php>. Accessed November 2017.
- Rasinski, T. (2017). Readers who struggle: Why many struggle and a modest proposal for improving their reading. *The Reading Teacher* 70(5), 519–524.
- Richards, J. C. (2010). Competence and performance in language teaching. *RELC Journal*, 41(2), 101–122.
- Robertson, K. (2009). Reading poetry with English language learners. <http://www.readingrockets.org/article/reading-poetry-english-language-learners>. Accessed November 2017.
- Storynory. (n.d.). <http://www.storynory.com/>. Accessed November 2017.
- Teacher Vision. (n.d.). *Reading aloud*. <https://www.teachervision.com/reading-aloud-0>. Accessed November 2017.
- Terhart, E. (2013). Teacher resistance against school reform: Reflecting an inconvenient truth. *School Leadership & Management*, 33(5), 486–500.
- Tyrell, J. (1997). Picture books and fantasy texts. In P. Falvey, & P. Kennedy (Eds.), *Learning language through literature: A sourcebook for teachers of English in Hong Kong*. Hong Kong University Press: Hong Kong.
- Vanderplank, R. (2009). Déjà vu? A decade of research on language laboratories, television and video in language learning. *Language Teaching*, 43(1), 1–37. <https://doi.org/10.1017/S0261444809990267>.
- Williams, R. (2014). *Why change management fails*. <https://www.psychologytoday.com/blog/wired-success/201411/why-change-management-fails>. Accessed November 2017.
- Worthy, J., Chamberlain, K., Peterson, K., Sharp, C., & Shih, P.-Y. (2005). The importance of read-aloud and dialogue in an era of narrowed curriculum: An examination of literature discussions in a second-grade classroom. *Literacy Research and Instruction*, 51(4), 308–322.
- Zairi, M., & Leonard, P. (1996). *Origins of benchmarking and its meaning*. Dordrecht: Springer Science and Business Media.
- Zakay, D. (1993). The impact of time perception processes on decision making under time stress. In O. Svenson & A. J. Maule (Eds.), *Time pressure and stress in human judgment and decision making*. N.Y: Plenum Press.

Peter Falvey is a teacher educator, formerly a Head of Department in the Faculty of Education, The University of Hong Kong. His main publication and research interests are in language assessment, first and second language writing methodology, and text linguistics.

David Coniam is Chair Professor and Head of Department of the Department of Curriculum and Instruction in the Faculty of Education and Human Development at The Education University of Hong Kong, where he is a teacher educator, working with teachers in Hong Kong primary and secondary schools. His main publication and research interests are in language assessment, language teaching methodology and computer-assisted language learning.

Index

A

- Advisory Committee on Teacher Education and Qualifications (ACTEQ), 21, 47, 162, 176, 240
- American Association for Counselling and Development (AACD), 30
- American Council on Teaching of Foreign Languages (ACTFL), 7
- American Educational Research Association (AERA), 31, 150
- American Psychological Association (APA), 150
- Analytically-marked, 32, 33, 108, 125, 126, 132, 311
- Angoff method, 32, 126, 254, 313–317, 319
- Assessment instruments, 4, 17, 32, 33, 48, 49, 54, 58, 59, 70, 88, 105, 125, 128, 131, 135
- Association of Language Testers in Europe (ALTE), 150
- Audiovisual, 407

B

- Battery of tests, 28, 32, 49, 54, 55, 57, 94, 118, 120
- Benchmark assessment, 7, 15, 41, 54, 105, 109, 138, 159, 410
- Benchmark initiative, 34, 59, 67, 68, 140, 400, 404
- Benchmark levels, 110, 114, 125, 140, 240
- Benchmark subject committee, 153
- Benchmark test, 32, 96, 117, 135, 136, 161, 240, 391

C

- Cambridge Examination in English for Language Teachers (CEELT), 54, 151
- Cambridge Proficiency in English Test (CPE), 151
- Canadian language benchmarks, 5, 14, 15, 18
- Chi square, 360, 364
- Chinese as a Medium of Instruction (CMI), 42, 43, 358
- Classical Test Theory (CTT), 140–142
- Classroom Language Assessment (CLA), 14, 22, 30, 33, 54, 56, 59, 66, 67, 69, 88, 97, 99, 100, 102, 108, 109, 115, 121, 128, 129, 170, 171, 175, 207–209, 222, 231, 247, 248, 250, 258, 286, 291, 294, 295, 306, 364, 381, 384, 391, 408, 411, 413
- Cloze test, 88, 94, 116, 118, 126, 127, 134, 138
- COBUILD, 4
- Cognitive abilities, 179–185, 192, 194, 196, 203, 245
- Common European Framework of Reference (CEFR), 313
- Complex structures, 77–82, 84, 234
- Consultancy study, 47, 48, 53, 58, 59, 63, 64, 66–68, 102, 106
- Consultancy team, 47–49, 53, 56, 58–60, 62, 64, 89, 150, 152, 153
- Content knowledge, 16, 17, 19, 21, 34, 49, 58, 176, 210, 213, 247, 280, 285, 411
- Course providers, 28, 60, 159, 160, 163–165, 171, 174, 175, 179, 292, 293, 399
- Credentialing teachers of English to speakers of other languages (ELSO), 17

- Criterion-referenced, 13–16, 29, 32, 33, 49, 116, 125, 131, 138, 207, 208, 318, 407
- Criterion-referenced assessment, 13, 31, 32, 60, 148, 407
- Criterion-referenced tests, 32, 88, 108, 125, 126, 138, 152
- Cut scores, 33, 60, 125, 138, 140, 150, 244, 245, 253, 254, 267, 268, 272, 279, 281–283, 308, 311–319, 400
- D**
- Data analysis, 144, 260, 261, 285, 371, 374, 393
- Data collection, 29, 72, 141, 350, 371, 374
- Degree courses, 41, 162, 358, 386, 393
- Discourse and use, 81–85, 169, 233
- E**
- Education and Manpower Bureau (EMB), 47, 70, 400
- Education Bureau (EDB), 8, 22, 44, 102, 164, 250, 252, 257, 258, 294, 295, 307, 314, 315, 325, 326, 357, 365, 393
- Education Commission, 20, 21, 41, 42, 44, 47, 239, 327, 400, 414
- Education Department (ED), 43, 49, 69, 97, 105, 106, 160, 239, 242, 244, 250, 357
- Education system, 6, 34, 37–39, 183, 252, 337, 338
- Educational change, 253, 320, 323, 324, 328–330, 338, 402
- Eigenvalues, 353
- English as a Foreign Language (EFL), 18
- English as a Medium of Instruction (EMI), 42, 358, 379
- English as a Second Language (ESL), 5, 14, 18
- English language, 3, 17, 19–21, 28, 39, 47–50, 53, 56, 58, 62, 66, 87–90, 92, 96, 98, 100, 102, 105, 108, 110, 111, 126, 133, 136, 147, 150, 162, 163, 175–177, 195, 212, 233, 234, 240, 242, 248, 249, 252, 253, 257, 258, 260, 271, 296, 305, 307, 311, 323–325, 328, 331, 350, 357, 358, 363, 365, 372, 373, 376, 377, 380, 381, 386, 387, 390, 392, 393, 400, 404–406, 409, 411
- English language benchmark, 21, 88, 90, 111, 135, 150, 152, 407
- English Language Benchmark Subject Committee (ELBSC), 34, 68, 70, 87, 125, 150, 328
- English language education, 7, 19, 20, 39, 251, 323, 376, 411
- English language heads of department, 371–373, 388–392
- English Language Learner (ELL), 17
- English language proficiency, 16, 19, 20, 43, 159, 161–163, 172, 175, 177, 351, 353, 366, 367, 377, 379, 381, 386, 388, 389
- English language standards, 6, 161, 212, 252, 349, 359, 364, 375, 378, 379, 383, 386–388, 392, 393, 405, 411, 412
- English language teachers, 3–6, 17, 19, 20, 22, 29, 40, 48, 49, 53, 56, 58, 59, 66, 89, 94, 95, 106, 118, 123, 135, 136, 161–164, 166, 175, 176, 199, 200, 204, 208, 215, 239, 241, 242, 248, 350, 351, 353, 357, 365, 372, 375, 376, 380, 388, 389, 391–393, 409, 411
- English language teaching, 19, 20, 22, 50–52, 59, 61–63, 67, 71, 94, 98, 100, 107, 136, 161–164, 175, 177, 200, 202, 204, 224, 245, 251, 335, 354, 355, 359, 360, 372, 392, 409, 412
- English major, 378, 382, 383, 386, 387, 404
- English proficiency, 4, 19, 20, 22, 126, 161, 170, 248, 319, 340, 354, 355, 360–363, 377, 379, 385
- Enhancement courses, 48, 160, 163, 164, 166, 170, 171
- Error correction, 168, 171, 174, 209–215, 279, 280, 297, 298, 329, 378, 386
- Exemptions, 33, 48, 70, 139, 354, 355, 360–364, 374, 383, 393, 404, 405, 412
- Expert judgement, 125, 126, 128, 138, 140, 244, 245, 266, 268, 272, 282, 283, 313, 315, 317, 318
- Explaining errors, 172, 209, 210, 214, 215, 246, 247, 279, 295, 387
- Explaining language matters to peers, 66, 91, 98, 99, 112, 113, 208, 216, 218, 221, 233, 247, 249, 250, 288, 337
- Expository writing, 76–78, 92, 93, 98, 115, 209, 210, 214, 246, 295, 297, 298
- F**
- Fair Average, 275, 276, 278, 289, 290, 294, 308, 355
- First Certificate in English (FCE), 151
- Focus group, 259, 263, 264, 273, 279, 280, 291
- G**
- Graduate Management Admission Test (GMAT), 5
- Graduating Students' Language Proficiency Assessment (GSLPA), 21

- Grammatical errors, 76–78, 80, 82, 84, 85, 213, 234, 250, 307, 380
- Grammatical knowledge, 180, 194, 213, 263, 380
- Grammatical structures, 76, 77, 79–84, 232, 234, 248, 304, 306, 307, 380
- Group discussion, 56, 216, 218, 221, 259, 291, 318, 382, 385
- Group interaction, 56, 58, 113, 216, 247, 285, 286, 288, 297, 298
- H**
- Head of department, 71, 251, 357, 372, 373, 382, 386, 388–390, 404, 411
- High-stakes assessment, 3–5, 11, 27–29, 31, 32, 239, 257, 323, 329, 400, 410
- HKSAR Government, 22, 28, 32–34, 43, 44, 68, 70, 102, 159, 162, 176, 242, 258, 311, 315, 326, 360, 371, 383, 403
- Hong Kong Advanced Level Examination (HKALE), 38, 39
- Hong Kong Advanced Supplementary Level Examination (HKASLE), 39
- Hong Kong Certificate of Education (HKCE), 38, 39
- Hong Kong Diploma of Secondary Education (HKDSE), 38, 39, 311
- Hong Kong Examinations and Assessment Authority (HKEAA), 22, 44, 87, 252, 259, 307, 350, 351
- I**
- Infit mean square, 110, 113, 147, 275, 289, 290, 355, 356
- In-service teachers, 41, 53, 68, 139, 159, 160, 163, 166, 224
- Interacting with peers, 66, 69, 79, 80, 91, 98, 99, 112, 113, 208, 216, 218, 221, 232, 247, 249, 289, 305
- Interaction with students, 66, 81–85, 306
- International English Language Testing System (IELTS), 5, 387, 388, 411
- Intonation patterns, 227
- Item difficulty, 127, 140, 142, 146, 254, 274, 314, 318
- Item Response Theory (IRT), 146
- L**
- Language awareness, 18, 28, 55, 57, 167, 168, 176, 217, 328, 332, 380
- Language learning, 13, 18, 100, 204, 271, 407
- Language of instruction, 38, 83–85, 90, 97, 99, 108, 109, 161, 208, 222, 224–226, 228, 231, 233, 247, 292, 293, 306
- Language of interaction, 66, 83–85, 89, 90, 97, 99, 108, 109, 208, 222, 224, 225, 228, 231, 233, 247, 292, 293, 306
- Language of presentation, 66, 81–84, 89, 233, 382
- Language problems, 58, 79–83, 100, 233, 247
- Language proficiency, 7, 14, 16, 17, 21, 33, 44, 148, 151, 159–164, 170–172, 175–177, 208–210, 216, 222, 225, 231, 233, 234, 239, 240, 247, 248, 252, 253, 257, 279, 280, 285, 286, 296, 305, 323, 335, 341, 365, 375, 377–384, 386–392, 400, 408, 410, 411
- Language Proficiency Assessment for Language Teachers of English (LPATE), 16, 22, 41, 140, 149, 159, 160, 164, 166, 167, 169, 171, 175, 181, 193, 203, 207, 209, 210, 214, 218, 222, 239, 241, 242, 244, 248, 250, 252, 253, 259, 262, 269, 275, 280, 285, 293, 294, 305, 311, 314, 317, 320, 323, 325, 327, 328, 330, 333, 337, 351, 358, 364, 375, 378, 381, 383–385, 387, 388, 390, 392, 393, 400, 404, 407, 411–413
- Language proficiency requirement, 159, 160, 164, 167, 171, 176, 177, 204, 241, 244, 262, 276, 294, 343
- Language proficiency standards, 47, 160, 239, 240, 366, 400
- Language Proficiency Test for Teachers (LPTT), 7
- Language standards, 3, 5, 7, 16, 20, 21, 59, 172, 179, 193, 330, 374, 377, 386, 392, 394, 412
- Language teachers, 7, 16–18, 28, 29, 42, 47, 49, 160, 162, 192, 240, 246, 252, 253, 264, 279, 327, 329, 331, 333, 341, 342, 377, 379, 381, 383, 386, 392, 399, 400, 410
- Language teaching purposes, 47, 50, 251, 354, 355, 360, 400, 405
- Lexical accuracy, 274
- Lexical accuracy and range, 273–275, 278, 288, 289, 292, 293, 304–306
- Linguistic skills and knowledge, 179–183, 192, 196, 203, 245
- Listening comprehension, 55, 57, 126, 180, 181, 195, 197, 271, 272
- Listening skills, 61, 94, 165, 174, 193, 195–197, 200, 202, 203
- Listening tasks, 100, 193, 196–200, 202
- Listening Test, 30, 63, 88, 94, 95, 108, 118, 119, 122, 123, 126, 127, 137, 169, 193, 246, 248, 252, 261, 269–272, 295, 407

- Lower secondary, 22, 48, 49, 53, 58, 62–66, 96, 105, 107, 111, 113, 136, 248, 331
- M**
- Macro listening skills, 193, 197, 199, 200, 202
- Many Facet Rasch Analysis, 108, 110, 112, 115, 123, 146, 147, 244, 260, 273, 274, 289
- Medium of instruction, 19, 42–44, 47, 51, 95, 212, 213, 239, 240, 259, 327, 328, 358, 360–363, 365, 381, 400, 403, 405, 406, 414
- Minimum standards, 49, 50, 59, 251, 365
- Monolingual, 409
- Morphosyntactic, 246
- Multimodality, 407
- Multiple choice, 12, 18, 22, 30, 63, 93, 94, 97, 99–101, 108, 116, 118, 119, 126, 136, 138, 152, 245, 246, 262, 263, 295, 307, 313
- N**
- National Council on Measurement in Education (NCME), 30, 150
- O**
- Oral interaction, 58, 65, 66, 79, 80, 91, 101, 112
- Oral Proficiency Interview (OPI), 7, 14
- Organisation and coherence, 62, 76–78, 93, 98, 99, 114, 115, 208–211, 216, 246, 273–275, 278, 304
- Organisation and cohesion, 65, 91, 112, 113, 208, 221, 232, 247, 249, 273, 274, 289, 305
- P**
- Paper and pencil tests, 6, 60, 89, 101, 102, 138, 139, 313
- Pedagogical knowledge, 17, 19, 161, 286, 376, 383, 391
- People's Republic of China (PRC), 20
- Permitted teacher, 40
- Pilot benchmark assessment, 96, 102
- Pilot Benchmark Assessment (English) (PBAE), 21, 34, 87, 102, 105, 151
- Point biserial correlation, 117–119, 144
- Post Graduate Certificate of Education (PGCE), 41, 136, 152
- Practitioner, 337
- Preliminary English Test (PET), 151
- Pre-service teachers, 53, 88, 139, 228, 262, 270
- Pretest, 94
- Pretesting, 94, 139
- Primary school, 38, 339, 352, 363, 377, 379
- Primary school teachers, 42, 51, 53, 107, 133, 351, 363–365, 373, 404
- Professional development, 14, 159, 162, 166, 175, 199, 207, 210, 264, 336
- Professional Teachers' Union (PTU), 51, 250, 252, 403
- Professional training, 14, 21, 40, 42, 106, 107, 133, 134, 162, 163, 239, 380, 386, 392, 404
- Proficiency assessment, 159, 177, 225, 233, 234, 240, 252, 257, 296, 305, 323, 331, 400
- Pronunciation stress and intonation, 65, 66, 90, 91, 97–99, 108, 109, 112, 113, 208, 209, 218, 221, 222, 224, 225, 232, 233, 247, 248, 284, 289, 293, 306, 382
- Prose passage, 92, 137, 216, 247–249, 252, 264, 283, 284, 287, 295, 406
- Prototype benchmark, 22, 48, 49, 87, 102, 110, 129, 132, 134
- Q**
- Qualified teacher status, 40
- Qualitative data, 120, 141, 371, 374, 393
- R**
- Rasch model, 128, 146, 147, 265, 267, 274, 282, 295, 318, 354
- Rating scales, 14, 15, 60, 61, 92, 147, 148, 208
- Raw score, 147, 260, 266, 271, 276, 282, 332, 407
- Reading and listening tests, 12, 33, 67, 87, 125, 179, 181, 245, 280–282, 295, 311
- Reading comprehension, 55, 57, 62, 101, 108, 168, 180, 181, 186–191, 245, 262–265, 269, 295
- Reading module, 168, 172, 182, 189, 192
- Reading process, 180, 186–189, 192
- Reading skills, 94, 180–182, 186, 187, 189, 203, 265
- Reading Test, 62, 63, 88, 94, 108, 116–118, 121, 122, 126, 127, 131, 137, 138, 245, 246, 248, 260, 262–270, 272, 295
- Registered teacher, 40
- Relevant degree, 8, 22, 51, 58, 133, 135, 358, 360, 365, 373, 386, 388
- Royal Society of Arts (RSA), 18
- S**
- Scales and descriptors, 15, 31, 56, 61, 62, 66, 67, 88, 89, 91–93, 108, 112, 114, 150, 169, 207, 208, 210, 222, 224, 225, 258, 261, 273–276, 279, 281, 286–289,

- 291–295, 299, 304, 305, 312, 358, 378, 407, 412
- Scholastic Assessment Test (SAT), 5
- SCMP, 340–343
- Secondary school, 38, 40, 55, 59, 96, 101, 136, 159, 202, 212, 220, 221, 315, 351, 352, 361–364, 379, 408
- Secondary school teachers, 21, 41, 42, 51, 53, 107, 133, 135, 176, 364, 365, 373
- Sentence stress and intonation patterns, 79–84, 233
- Signalling devices, 81–85, 233
- Speaking Test, 12, 34, 61, 64, 66, 67, 88, 90–93, 108, 111–113, 121, 131, 137, 139, 169, 208, 216, 218, 222, 247–250, 252, 283–291, 295, 312, 364, 384, 387, 390, 413
- Standard Deviation (SD), 117–119, 142, 145, 147, 266, 271, 282, 359
- Standard Error of Measurement (SEM), 117–119, 145, 266, 271, 282
- Standard setting, 128, 149, 150, 152, 265, 266, 268, 272, 282, 295, 312, 313, 316, 317, 319
- Standing Committee on Language Education and Research (SCOLAR), 61
- Stereotypes, 329, 334
- Student composition, 34, 76–78, 93, 98, 100, 101, 114, 137, 295
- Subject-matter knowledge, 16, 18, 19, 49, 52, 53, 55, 58, 161, 213, 376, 380, 383, 391–393, 412
- Subject trained, 41, 253, 327, 386
- Subjective, 316, 334
- Subtest, 57
- Subthemes, 390
- Syllabus specifications, 102, 225, 233, 234, 257–259, 263, 299
- Syntactic structures, 182, 186–188, 192
- T**
- Task completion, 62, 65, 66, 72, 76–78, 93, 99, 114, 115, 208, 210, 211, 246, 273–275, 278, 304
- Teacher certification, 5, 15, 17, 28, 48, 49, 400
- Teacher education, 15, 16, 18, 20, 37, 40, 41, 48, 59, 153, 211, 242, 253, 280, 307, 323, 324, 364, 406, 408, 410
- Teaching experience, 48, 153, 162, 200, 351, 358, 361–365, 372, 373
- Teaching Knowledge Test (TKT), 18
- Teaching profession, 7, 8, 18, 40, 41, 160, 252, 358, 392, 393, 412
- Teaching purposes, 355
- Teaching qualification, 22, 49, 50, 59, 68, 69, 251
- Teaching skills, 174, 329, 333, 343, 382, 391, 411
- Technique, 229
- Tertiary institutions, 4–6, 41, 68, 152, 160, 163, 242, 307, 315, 406
- TESL, 14
- Test component, 67, 105, 120, 131, 132, 262, 358
- Test developers, 13, 28, 32, 153, 249, 252, 258, 259, 263, 265, 269, 270, 279, 285, 291, 294
- Test equating, 125, 129, 244, 295, 314, 316, 318
- Test for English Major students, 20
- Test items, 11, 12, 17, 94, 128, 136, 146, 179, 244, 265, 271, 281, 282, 313
- Test of English as a Foreign Language (TOEFL), 19
- Test papers, 18, 31, 245, 246, 258, 265, 307, 315
- Test results, 14, 147, 208, 264, 266, 271, 274, 276, 282, 289, 330
- Test takers, 13, 21, 28, 30–32, 34, 55, 61, 64, 67, 68, 92, 94, 95, 105, 109, 112, 113, 115–123, 131–136, 138, 146–148, 244, 245, 260–262, 265, 266, 268, 271, 273–276, 278, 281, 282, 286, 288–290, 313
- Test types, 32, 48, 53–55, 57, 88–90, 92, 108, 111, 120, 128, 129, 134, 135
- The Chinese University of Hong Kong (CUHK), 40, 41, 55, 71, 152, 160, 163, 165, 166, 171, 176, 179, 350, 386, 400, 411
- The Education University of Hong Kong (EdUHK), 40, 163, 262, 350, 352, 411
- The Hong Kong Institute of Education (HKIEd), 40, 152, 163, 165, 175, 262, 291, 351
- The University of Hong Kong (HKU), 40, 41, 71, 152, 261, 386
- TOEIC, 19

Training courses, 20, 40, 41, 159, 160,
162–166, 171, 172, 175, 179, 193–195,
199, 204, 210, 222, 241, 251, 336, 351
Transcription, 217
Translanguaging, 409, 410, 414

U

UGC, 399
Undergraduates, 6, 379
Unidimensionality, 128
University of Cambridge Local Examinations
Syndicate (UCLES), 18
Use of English (UE), 43, 95, 118, 250

V

Videoing, 56, 58
Volunteering, 106, 135

W

Writing module, 168, 169, 207, 210, 211,
213–215, 218, 224
Writing task, 62, 93, 98, 100, 101, 210, 212,
246, 273
Writing Test, 34, 55, 61, 62, 67, 88, 92, 93,
108, 113–116, 122, 131, 134, 137, 147,
170, 172, 208–211, 214, 245–247,
272–274, 279–282, 285, 290, 295, 311,
312, 332, 333, 384, 407
Written Proficiency Test (WPT), 7