Lessons and Legacy: A Tribute to Professor Cyril J Weir (1950–2018)

Also in this series:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Language Assessment for Multilingualism: Proceedings of the ALTE Paris Conference, April 2014
Edited by Coreen Docherty and Fiona Barker

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

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

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

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

Applying the socio-cognitive model to the BioMedical Admissions Test
Edited by Kevin Y F Cheung, Sarah McElwee and Joanne Emery

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

Lessons and Legacy: A Tribute to Professor Cyril J Weir (1950–2018)

Edited by

Lynda Taylor
Centre for Research in English Language Learning and Assessment, Bedfordshire

and

Nick Saville
Cambridge Assessment English

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

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

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

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

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

Cambridge University Press is part of the University of Cambridge.

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

www.cambridge.org
Information on this title: www.cambridge.org/9781108745734

First published 2020

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

Printed in XXXX by XXXX

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

ISBN 978-1-108-74573-4

Professor Cyril James Weir

1950–2018

Contents

Measures of Esteem

Appendices

Acknowledgements

In bringing this volume to fruition, we are indebted to a significant number of Cyril Weir's friends and colleagues who generously offered to share their memories of and reflections on working with him in different contexts and at different times during his long career in English language teaching, learning and assessment. Their involvement is all the more appreciated given the ambitious timeframe for this project, which arose from our heartfelt desire to publish a tribute volume as quickly as possible following Cyril's untimely death in September 2018.

Doing justice to the achievements of an academic and scholar like Cyril Weir over such a long and distinguished professional career is not an easy task, but the contributions of all the authors mentioned below will hopefully provide a sense of the considerable impact he had on individuals and groups as well as of the legacy he left to our field as a whole.

Special thanks must go to those who provided the extended papers that make up the core of this volume in Chapters 1 to 8: Vivien Berry, Barry O'Sullivan, John Field, Yan Jin, Sathena Chan, Nicola Latimer, Fumiyo Nakatsuhara, Jamie Dunlea, Guoxing Yu, Tony Clark, Lynda Taylor and Anthony Green.

Other colleagues – some from the early years of Cyril's career, others from later years – supplemented the longer papers by providing shorter Measures of Esteem. These contributions bring rich personal as well as professional insights into what it meant to have Cyril as both colleague and friend, and we are most grateful for the recollections of Roger Hawkey, Eddie Williams, Jon Roberts, Hanan Khalifa, Jessica Wu and Lynda Taylor.

A number of other language testing colleagues around the world who knew and worked with Cyril over the years shared their memories and appreciation of Cyril in conversations and email exchanges with us, and these too have fed into this publication more informally. Though they are too numerous to mention by name we thank them all most warmly for their contribution to this volume.

Appreciation must also go to John Savage, Publications Assistant at Cambridge Assessment English, who assembled and compiled the appendices for the volume. His careful and supportive management of the publication project enabled us to take this edited volume from initial concept through to printed volume in little more than 12 months – no mean feat!

To all of the above, and to any others we have failed to mention, we extend our sincere thanks and appreciation.

Preface

In editing this volume as a tribute to Cyril Weir, we have attempted to put together a narrative that tells a story about his work and the legacy he leaves behind. It seemed to us to be an appropriate genre to remember him by given that Cyril was a historian at heart and he used his personal, professional and institutional connections in piecing together his own understandings of the past.

In choosing the title *Lessons and Legacy*, the intention is to tell the story through the voices of the people that he influenced and who played a part in his long career spanning more than 40 years. The Lessons are represented in a series of eight chapters that reflect his academic contributions to the field of language assessment. As a complement and extension to this, the Measures of Esteem section that follows brings in personal experiences and lessons learned in collaborating with Cyril, both as a colleague and as a friend. These unique reflections bring insights into his character and his working practice that can be instructive to others who aspire to work collaboratively, across continents and across disciplines.

An appreciation of the Legacy he leaves behind is voiced in part by those who are now taking forward the lessons learned from Cyril in their own careers. We are pleased that the volume has contributions from several leading academics who studied under Cyril or worked alongside him in the teams that he led. They have now become thought leaders in their own right and it seems from their papers that his legacy lives on in a new generation of students and practitioners in our field.

The current Series Editors of Studies in Language Testing (SiLT), Nick Saville and Lynda Taylor, thought it fitting for volume 50 to be dedicated to Cyril's memory, and in recognition of his contribution to the SiLT series as both author and joint Series Editor, first with the founder of the series, Michael Milanovic, and then with Nick Saville from 2014.

The narrative begins during the early days of Cyril's career that take us back to his doctoral studies at the Institute of Education (London), and very importantly to the University of Reading in 1986. Reading plays a significant part in the story – as testified in many of the papers – but also because the suggestion to produce this tribute volume was made at the university during a memorial lunch for Cyril organised by friends and colleagues in October 2018 soon after his funeral.

The original conceptualisation for this book began to crystallise on hearing the eulogies and personal reflections from his colleagues and friends

at that Reading event. A month later at the Language Testing Forum (LTF) 2018, fittingly hosted by the Centre for Research in English Language Learning and Assessment, Bedfordshire (CRELLA) in Putteridge Bury, the commitment to produce the volume was announced. Further tributes and reflective talks were given during LTF and the editors were able to draw on those contributions in finalising this volume. It is testimony to the esteem in which Cyril was held that all the contributors agreed to meet very tight deadlines in getting the volume finished in less than a year. We are grateful to them and acknowledge the efforts that they have made to bring the volume to fruition.

Connections with Reading are also important for Cambridge Assessment English (henceforth Cambridge English) in a number of other ways, and the value of long-term personal and institutional collaborations emerge as part of this narrative. By the 1980s, the Department of Linguistics and the Centre for Applied Language Studies (CALS) at the University of Reading had gained a world-leading reputation and were in the forefront of research in Applied Linguistics and TEFL. Cyril joined CALS in 1986 during that vibrant period and bolstered up an already flourishing Testing and Evaluation Unit (TEU) alongside Arthur Hughes and Don Porter. The relationship between Reading and Cambridge English was established in those early days through personal connections and fruitfully evolved during the 1990s. Part of this story is told by Lynda Taylor and Anthony Green in Chapter 8 of this volume.

The possibility of building up the relationship with Reading was enabled in Cambridge English when the first Evaluation Unit was set up in 1989. It became a part of the newly established EFL Department within UCLES (as it was then) and was led by Michael Milanovic with Nick Saville (both Reading graduates) in order to develop the research and validation capacity. Informal collaboration continued over the following years but before the end of the decade a Memorandum of Understanding (MOU) had been signed between Cambridge English and the TEU led by Cyril. The stated objective was to collaborate on projects of mutual interest commissioned by Cambridge English in order to extend the capacity to carry out research in support of the Cambridge English examinations. These projects, although led by Cyril and his colleagues in Reading, were carried out in conjunction with researchers and practitioners in Cambridge English. Of particular note is the collaborative work conducted in the late 1990s after Barry O'Sullivan joined the TEU whilst he was completing his PhD. Several joint publications and presentations emerged, especially in the area of speaking test validation (e.g. O'Sullivan, Weir and Saville 2002) and Barry's paper in this volume (Chapter 2) is based on work commissioned by Cambridge English at that time. This collaboration between the two institutions underpinned the socio-cognitive framework that was first adopted by Cambridge in 2003–04.

In many ways, the evolving socio-cognitive model and its practical application in test development and validation came to fruition after Cyril left Reading. The well-known version of his model was published in 2005 and proved to be very influential. Arguably it forms the centrepiece of what he himself considered his most productive period, from 2006 onwards, following the establishment of CRELLA. It is certainly a central theme in the story being told in this volume, and the legacy of this work lives on in CRELLA, as testified by his colleagues who have contributed to this narrative.

The connection between Cambridge English and Cyril continued to flourish in the CRELLA period and the evolving socio-cognitive model became central to the validation of Cambridge English examinations. This is described and exemplified in the four volumes in the SiLT series that Cyril contributed to as author and editor – *Examining Writing*, *Examining Reading*, *Examining Speaking*, and *Examining Listening* (respectively, Shaw and Weir 2007, Khalifa and Weir 2009, Taylor (Ed) 2011, Geranpayeh and Taylor (Eds) 2013).

The origins of Cyril's preoccupation with certain elements of his model can be traced back to the very beginning of his career and to the formative period in his life when he was working on his doctoral studies. There is a strand of thinking and an intertwining of personal relationships in this volume that can be teased apart to show this. In his early work on English for Specific Purposes (ESP) and during the development of his methodology for analysing the communicative needs of students, he showed a particular concern for the practical realities of learning and assessment. The origins of the *contextual* aspects of his model can be located there. Cyril's research methods rejected armchair theorising that he thought would lead to impractical solutions; rather, an empirically based understanding of the educational context was central to his thinking. He concluded that the tasks for assessing academic English should relate to the actual *activities (or operations), conditions* and *contexts* under which they are performed in the target setting.

Friendships formed at that time also proved to be long-lasting and influential and we hope we have captured this part of the story in the Measures of Esteem section. Cyril acknowledged the work of Roger Hawkey, a fellow doctoral student who was working on the cognitive dimensions of a similar topic, and their academic collaboration was renewed many years later when Roger began working as a consultant for Cambridge English (2000 onwards), and subsequently when he joined CRELLA as a visiting professor (see the first of the volume's Measures of Esteem, written by Roger).

The Cambridge English connection crops up in several other strands of the narrative including Cyril's passion for history and his historical research; his insightful editorial skills; and his long-standing interest in the assessment of academic reading. These three strands come together in Cyril's contributions

to the SiLT series, and through his longstanding collaboration with Lynda Taylor and Nick Saville in that context.

Michael Milanovic invited Cyril to join the SiLT editorial team when the series was already well established (14 volumes), soon after they had worked together on editing SiLT 15, the volume that documents the history and revisions of the Cambridge English Certificate of Proficiency in English (CPE) (Weir and Milanovic (Eds) 2003). During that process, Cyril found his way into the archives of Cambridge Assessment and he was immediately in his element there. Following on from that experience, Cyril returned many times to the archives and also sought out witnesses and voices from the past in preparing the centennial volume *Measured Constructs: A History of Cambridge English Language Examinations 1913–2012* (SiLT 37, 2013), co-authored with Ivana Vidaković and Evelina Galaczi. He was particularly proud of this volume and remained very supportive of this strand of publishing within SiLT, including the volumes that document the history of specific examinations using archival data and first-hand evidence from stakeholders, such as *A Modular Approach to Testing English Language Skills: The development of the Certificates in English Language Skills (CELS) examinations* (SiLT 16, Hawkey 2004) and *Assessing Academic English: Testing English proficiency, 1950–1989 – the IELTS solution* (SiLT 23, Davies 2008).

Cyril's skill as an editor included his ability to read fast and efficiently, as well as his insightful judgements about the merits of a manuscript. He insisted on high standards and in maintaining editorial integrity, but he was always prepared to work with authors to improve their proposals for inclusion in the series. He readily provided mentorship and support for early career researchers submitting their doctoral theses for publication. His constructive approach also meant that he was the ideal editor for conference proceedings or collected papers where diverse contributions needed to be collated and edited with sensitivity.

It is fitting that part of his legacy includes his last volume as an author in the SiLT series. He was actively working on the manuscript of volume 51 with his co-author Sathena Chan at the time of his illness in 2018. Entitled *Research and Practice in Assessing Academic Reading: The Case of IELTS*, the volume was successfully published posthumously and provides an appropriate tribute to Cyril's work in the field of academic reading (Weir and Chan 2019).

The nature of the academic reading construct had been a key concern for him throughout his career and he was constantly looking for innovative ways to extend the range of task types that can be operationalised in assessing academic reading, e.g. expeditious reading tasks, reading-into-writing tasks, etc. This final volume provided him the opportunity to reflect on over four decades of research into the theory and practice and he was able to leave us with some insights about the future of assessment in this domain. His

collaboration with Sathena, as PhD supervisor and co-author of the volume, also attests to Cyril's strength as a mentor. This quality is recognised as an important part of his legacy and is endorsed by several other contributors to this volume.

In the next section, the overarching narrative is outlined in more detail. The eight main chapters are summarised and some key points in the Measures of Esteem are highlighted for the convenience of prospective readers.

Chapter 1 takes us back to the very beginnings of Cyril Weir's career as a professional language tester and to his doctoral studies referred to above. In her contribution, Vivien Berry describes the development of empirical methods of analysis for determining communicative needs. She examines attitudes and beliefs to designing and validating test items in each of the different eras of language testing from the 1970s to the present day. Her chapter recalls a number of shifts in approach over that period: from *a posteriori* test analysis to *a priori* test design; from discrete-point general English proficiency tests to ESP-focused tests involving more communicative tasks; and from discipline-specific tests to single tests of overall communicative competence. Vivien exemplifies the application of empirical needs analysis from the 1980s onwards by reference to the Test in English for Academic (later Educational) Purposes (TEAP/TEEP), the Occupational English Test (OET) and the English Language Testing Service (ELTS), which later became the International English Language Testing System (IELTS). Her contribution reflects upon the enduring relevance of Cyril Weir's early PhD study on empirical needs analysis, an important piece of work which not only helped to underpin test design, development and validation work undertaken during the 1980s and 1990s for tests such as TEAP/TEEP, OET and ELTS/IELTS, but which still serves today as an essential research methodology for the creation of tests that are appropriate for specific populations.

Chapter 2 is a much-abbreviated version of a historical report on testing spoken language commissioned from Barry O'Sullivan and Cyril Weir by Cambridge English in 2002. When it was first produced, their report offered an overview of speaking assessment and associated research issues, which was designed to inform the Cambridge English test development and validation agenda at that time. Consistent with the long commitment of Cambridge English to the direct assessment of speaking and writing ability, a fruitful synergy had been established with Barry and Cyril by the mid-1990s, as noted above, and when the Performance Testing Unit was established in 1999 by Peter Hargreaves under the leadership of Lynda Taylor, a stronger focus on research was already emerging. Their report presented a valuable framework through which previous research could be evaluated and future research in the area might be formulated. The reason for including an abbreviated version of that early report in this edited volume is that it helps to show the

genesis of the socio-cognitive framework which has become so instrumental in recent years for many parts of the language testing community and which is closely, though not exclusively, associated with Cyril. A socio-cognitive approach to language testing and assessment, as outlined in the socio-cognitive framework, has helped the field to reconceptualise and enhance previous views of construct validity, allowing much greater attention to be paid to the category of theory-based, cognitively-related validity alongside the more well-established categories of content and context validity.

In Chapter 3, John Field traces some of Cyril's innovative thinking in this area as he explored the growing body of research concerned with the human command of cognitive processes leading to behavioural performance, rather than just knowledge of information and the individual's ability to declare it. John reflects on how Cyril's enquiring mind and his willingness to make interdisciplinary connections with a parallel field of enquiry (i.e. the nature of expertise and expert performance) was a hallmark of his academic openness and rigour. Cyril always recognised that one needs to be willing to draw on and take due account of the views and expertise of specialists in other domains. John's chapter provides an accessible and helpful background to the notion of cognitive validity and its recent role in second language testing, especially in relation to the skill of reading, which was a key area of interest for Cyril throughout his life.

Chapter 4 focuses attention on the other core validity category prioritised by Cyril within a socio-cognitive approach – context validity. Context validity can be defined as the characteristics of any test, embracing aspects of the input and the task but also the circumstances under which the test is completed. In her contribution, Yan Jin demonstrates the value of Weir's 2005 socio-cognitive framework for operationalising the constructs of two locally developed English language tests in the Chinese context. She reviews the general notion of context validity and exemplifies the central role of contextually appropriate operations and conditions in construct operationalisation through a detailed analysis of the expeditious reading tasks in the Advanced English Reading Test (AERT) and the peer-to-peer discussion in the Spoken English Test component of the College English Test (CET-SET). Her chapter ends with reflections on the lessons learned from these two case studies and on the inherent challenges in extrapolating from test performances to performances in target language use situations. She also acknowledges Cyril Weir's considerable legacy in China with regard to the development of professional expertise and experience in the field of language assessment.

In Chapter 5, Sathena Chan and Nicola Latimer highlight Cyril's research into the nature of academic reading. Using one of his recent test development projects as an example, they describe how the construct of academic reading was operationalised in the local context of a British university. This was achieved by combining theoretical construct definition with empirical

analyses of students' reading patterns on the test using eye-tracking methodology. The authors discuss how Weir's extensive theoretical and empirical research into the nature of academic reading over 35 years fed into the development of the new university reading test, as well as a new method of analysing eye-tracking data in relation to different types of reading. Their chapter includes a series of engaging and illuminating personal reflections on what it was like to experience Cyril as PhD supervisor, research mentor and professional senior colleague.

Chapter 6, written by Fumiyo Nakatsuhara and Jamie Dunlea, focuses the spotlight back onto speaking assessment. They describe how the socio-cognitive framework guided two *a priori* validation studies for the speaking component of the Test of English for Academic Purposes (TEAP), a new admissions test for colleges and universities in Japan with components covering all four language skills, including a face-to-face speaking test. They discuss the effectiveness and value of the framework in underpinning test design for the TEAP and in gathering empirical evidence of the construct underlying a speaking test for the target context. Like Sathena and Nicola, the authors reflect upon Cyril's significant personal contribution to developing early career researchers in the field; and, like Yan Jin, they acknowledge his lasting contribution to extending expertise in the East Asian context through his close collaboration with the TEAP development team.

Chapter 7, written by Guoxing Yu and Tony Clark, continues the thread of construct operationalisation in testing practices, shifting the focus onto the assessment of reading and writing through the use of integrated reading-into-writing tasks. The authors explore Cyril's early thinking on the use of integrated tasks from his PhD research in the 1980s through to his publications in the 1990s on test design and development, and finally to his SiLT series volumes on assessing reading and writing in the 2000s. They highlight his concern over potentially 'muddied measurement' when assessing skills in combination, but also his advocacy of integrated reading-into-writing tasks as being both cognitively and contextually valid, especially in the context of English for Academic Purposes (EAP) assessment. Guoxing and Tony perceive a steady evolution in Cyril's views over time as he encountered new research findings and as he engaged in practical test development projects, reflecting his ongoing commitment to an evidence-based approach to language test development and validation.

In the final chapter, Lynda Taylor and Anthony Green reflect upon the role of academic institutions in language testing research and consultancy over the past 50 years, specifically the part played by university-based departments or research centres in developing theory and practice in the field of language testing and assessment. Chapter 8 acknowledges a part of the story of our professional field which has not received much attention and it offers an appreciation of the contribution of selected individuals, teams and

organisations within that story. The authors describe a range of academic institutional contexts, in the UK, the US, Australia and Canada, in which research has flourished at different times over recent decades, considering who the key players were and what the individuals and teams within them accomplished. Several of these were universities with which Cyril had a close personal connection as a postgraduate student, as junior lecturer, as Centre Director and as Professor. Lynda and Anthony reflect upon the significance and impact of such institutions and their legacy with regard to current theory and practice in language testing and assessment.

Contributors to the Measures of Esteem section were invited to offer something closely focused and personal in which they might describe an area they collaborated on with Cyril and what they had learned from that experience. The written contributions that were provided therefore contain personalised accounts, memories and reflections from individuals who knew Cyril well and who worked with him professionally in different contexts and at different times during his long and illustrious career. They supplement the longer papers in Chapters 1 to 8, providing professional and personal measures of esteem, and bringing to a wider readership rich, and sometimes amusing, insights into what it meant in practice to have Cyril as both colleague and friend.

Roger Hawkey recalls meeting Cyril in the late 1970s when they were both postgraduate students at the Institute of Education, London University. Their overlapping PhD interests in the English language skills and needs of international students led to a friendship which lasted a lifetime and saw extensive research collaboration, especially concerning the assessment of academic reading ability. In his reflection, Roger discusses the relevance of some innovative research into the academic reading construct completed in the early 2000s while both he and Cyril were working at CRELLA.

Eddie Williams first met Cyril in 1986 when they became colleagues at CALS in the University of Reading. Though they shared a common interest in the teaching and testing of reading skills, they held sharply differing views on the actual nature of reading ability and how it should be assessed. Eddie recalls having a vigorous debate with Cyril on the issues which ultimately led to a decision not to try and co-author a book on the subject. Happily, their personal and professional friendship was maintained through a shared love of international rugby and good beer!

Jon Roberts was also a colleague at the University of Reading in the late 1980s/1990s. He recalls collaborating with Cyril on several international education projects, including a project funded by the Overseas Development Administration (ODA) with secondary-level teachers of English in Nepal. Their collaboration in this area led to a co-authored volume in 1994 entitled *Evaluation in ELT*. In his reflection, Jon recalls the pleasure of undertaking an authorial apprenticeship with Cyril; he recalls the latter being

'particularly, one might say forensically, strong on coherence and relevance' but also someone who made book-writing fun.

Hanan Khalifa also reflects on her experience of Cyril's role in national assessment reform initiatives, from the perspective of her home country of Egypt between 1991 and 2003. She highlights Cyril's ability to bring to a project an international perspective while still remaining sensitive to and showing understanding of the local context, with its opportunities and its constraints. She also notes how his mentorship, guidance and example helped train and equip a cadre of well-qualified language testers for the Egyptian context.

From another part of the world, Jessica Wu reflects on the contributions of Cyril Weir to the General English Proficiency Test (GEPT) in Taiwan from 2001 onwards. Like Hanan, Jessica comments on how Cyril constantly sought a healthy balance between international standards/expectations and appropriate localisation in test development, especially with regard to the local learning context and curriculum. He encouraged a strong commitment to quality assurance and validation research in the Taiwanese language testing context and personally mentored several of the emerging academics in that context. Jessica notes that the volume *English Language Proficiency Testing in Asia: A New Paradigm Bridging Global and Local Contexts*, which Cyril had just completed co-editing with her and Lily I-Wen Su at the time of his death, testifies to his enthusiastic support for language testing in the region.

Lynda Taylor collaborated with Cyril for many years on the SiLT series. She also co-taught many of the Association of Language Testers in Europe (ALTE) courses on testing and assessment with him during the 2000s and was a colleague at CRELLA from 2011. In her contribution she shares memories of travelling with Cyril to different European cities as they lectured together on the ALTE training courses and what a good travel companion he proved to be. She also reflects on the confidence Cyril placed in his students and his colleagues, the collaborative research and publication opportunities he offered them, and particularly the encouragement and support he gave to female academics at the start of their career.

In 2014, Professor Cyril J Weir was awarded the Cambridge-ILTA Distinguished Achievement Award in recognition of his significant contribution to the field of language testing and assessment over many years. Cyril received his award at the 2014 Language Testing Research Colloquium (LTRC) conference in Amsterdam and in his invited lecture, he referenced the words written by 17th century mathematician, astronomer and physicist Sir Isaac Newton, in a letter to philosopher, architect and polymath Robert Hooke: *If I have seen further it is by standing on the shoulders of giants.* Cyril graciously acknowledged those who had gone before him in our field, 'giants' whose achievements had enabled him to extend his understanding

and develop his expertise in language testing and assessment. Many of us remember with gratitude Cyril's lessons as recounted in this volume and we are confident that his legacy will provide broad shoulders on which future generations of language testers can stand and look to the future.

Nick Saville
Lynda Taylor
December 2019

References

Davies, A (2008) *Assessing Academic English: Testing English Proficiency, 1950–1989 – The IELTS Solution*, Studies in Language Testing volume 23, Cambridge: UCLES/Cambridge University Press.

Geranpayeh, A and Taylor, L (Eds) (2013) *Examining Listening: Research and Practice in Assessing Second Language Listening*, Studies in Language Testing volume 35, Cambridge: UCLES/Cambridge University Press.

Hawkey, R (2004) *A Modular Approach to Testing English Language Skills: The Development of the Certificates in English Language Skills (CELS) Examinations*, Studies in Language Testing volume 16, Cambridge: UCLES/Cambridge University Press.

Khalifa, H and Weir, C J (2009) *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing volume 29, Cambridge: UCLES/Cambridge University Press.

O'Sullivan, B, Weir, C J and Saville, N (2002) Using observation checklists to validate speaking-test tasks, *Language Testing* 19 (1), 33–56.

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

Su, L I-W, Weir, C J and Wu, J R W (Eds) (2019) *English Language Proficiency Testing in Asia: A New Paradigm Bridging Global and Local Contexts*, London/New York: Routledge.

Taylor, L (Ed) (2011) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press.

Weir, C J and Chan, S (2019) *Research and Practice in Assessing Academic Reading: The Case of IELTS*, Studies in Language Testing volume 51, Cambridge: UCLES/Cambridge University Press.

Weir, C J and Milanovic, M (Eds) (2003) *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*, Studies in Language Testing volume 15, Cambridge: UCLES/Cambridge University Press.

Weir, C J and Roberts, J (1994) *Evaluation in ELT*, Oxford: Blackwell.

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

Notes on contributors

Vivien Berry was Senior Researcher in English Language Assessment at the British Council until January 2019. Major projects she completed with external colleagues include an investigation into the comparability of the construct of speaking in face-to-face and computer-mediated delivery of IELTS Speaking tests, and a multi-method investigation into language teachers' assessment literacy. She is currently co-editing a special issue of *Assessment in Education* exploring the use of innovative technology in oral language assessment and co-authoring a volume named *Assessing Speaking: Current and Future Perspectives* for the *British Council Monographs on Modern Language Testing* series.

Sathena Chan is a Senior Lecturer in Language Assessment at the Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire. Her research interests include integrated assessment (e.g. validation, task design, rating scale development), L2 language processing (reading, writing and summarising), and use of process-tracking technologies in language learning and assessment. She has conducted funded research projects and test development projects for leading examination boards in the UK and around the world. Her work has been published in peer-reviewed journals such as *Assessing Writing* and *System*. Her other recent publications include *Defining Integrated Reading-into-Writing Constructs: Evidence at the B2–C1 Interface* (English Profile Series volume 8, 2018) and *Research and Practice in Assessing Academic Reading: The Case of IELTS* (Studies in Language Testing volume 51, co-written with Cyril J Weir, 2019), both co-published by Cambridge Assessment English and Cambridge University Press.

Tony Clark is a Senior Research Manager at Cambridge Assessment English, managing research on the IELTS exam. His PhD focused on how Chinese and Japanese students learn to write in academic English, and how test preparation in their country of origin relates to their subsequent UK university experience. His thesis (funded by the Economic and Social Research Council (ESRC) and supervised by Dr Guoxing Yu and Dr Talia Isaacs) received a British Council Research Assessment Award in 2014. In 2015/2016, he was a recipient of the Newton Fund Scholarship (a grant to promote researcher mobility and encourage British-Chinese academic relations), and received

funding from the Worldwide Universities Network (WUN) and the ESRC to support overseas research trips. In 2016 he spent two months at the British Council in Tokyo and six months at Zhejiang University in China. He has contributed to research projects on language and admissions testing, language acquisition and test development, collaborating with Bristol University, Swansea University, Assessment Europe and the British Council throughout the MENA/Asia region and beyond.

Jamie Dunlea is a Senior Researcher and Manager of the Assessment Research Group at the British Council. Jamie works on a range of language test development and validation projects for assessment systems designed and developed by the British Council, as well as collaborating on projects with researchers and organisations internationally. Jamie has advised Ministries of Education and national agencies on large-scale assessment reform projects, overseen research for collaborative, international projects such as linking UK examinations to China's Standards of English, and is active in the language assessment research community. He joined the British Council in 2013, and was previously Chief Researcher at Eiken Foundation of Japan, a not-for-profit organisation which develops and administers EFL examinations in Japan. He has 25 years' of experience in EFL education, first as a teacher, then in test development and production and assessment research.

John Field is Reader in cognitive approaches to language learning at the Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire. His main area of expertise lies in second language listening, on which he has researched and written widely; his *Listening in the Language Classroom* (Cambridge University Press, 2008) is a standard work in the field. He has a background in psycholinguistics, which he taught for several years at the University of Reading, UK, at both undergraduate and postgraduate level. Research in this area has greatly shaped his thinking; and his most recent work has entailed applying psycholinguistic principles to the notion of cognitive validity in language testing. Before becoming an academic, he worked in many parts of the world as an ELT advisor, materials writer, curriculum designer and teacher trainer.

Anthony Green is Professor in Language Assessment and Director of the Centre for English Language Learning and Assessment (CRELLA), University of Bedfordshire and a former President of the International Language Testing Association (ILTA). He has consulted, presented and published on a wide range of language assessment issues around the world. He is the author of *Exploring Language Assessment and Testing* (Routledge, 2013), *Language Functions Revisited: Theoretical and Empirical Bases for Language*

Construct Definition Across the Ability Range (English Profile Studies volume 2, 2012) and *IELTS Washback in Context: Preparation for academic writing in higher education* (Studies in Language Testing volume 25, 2007), both co-published by Cambridge Assessment English and Cambridge University Press. His main research interests lie in the relationships between assessment, learning and teaching.

Roger Hawkey's postgraduate qualifications, research and professional career have been in English language teaching and assessment. He has held British Council posts in East and West Africa as well as London, and participated in further EL teaching and testing projects in Thailand, Europe and Latin America. The key focus of his most recent research projects has been on the impact of high-stakes English language exams, mainly through consultancies with Cambridge Assessment English and his Visiting Professorship at the Centre for Research in English Language Learning and Assessment (CRELLA) at the University of Bedfordshire. He has written four volumes in the Cambridge Assessment English/Cambridge University Press Studies in Language Testing (SiLT) series. These have been on: the First Certificate in English (FCE); the Certificate in Advanced English (CAE); the IELTS test and *Progetto Lingue 2000*; the Certificate in English Language Skills (CELS) exams; and, with Michael Milanovic, on the history of Cambridge English Qualifications in general.

Yan Jin is a Professor of Linguistics and Applied Linguistics at the School of Foreign Languages, Shanghai Jiao Tong University. Her research interests include the development and validation of large-scale, high-stakes language assessments. She is currently Co-President of the Asian Association for Language Assessment and Chair of the National College English Testing Committee of China. She is co-editor-in-chief of the Springer open-access journal *Language Testing in Asia* and is also on the editorial boards of *Language Testing*, *Language Assessment Quarterly*, *Classroom Discourse*, *International Journal of Computer-Assisted Language Learning*, and a number of academic journals published in mainland China.

Hanan Khalifa is a leading language testing and evaluation expert. Since 1993, she developed and validated national and international assessments, and led the alignment of locally produced curricula and examinations to international standards. She has presented and published extensively on various assessment topics. She has worked and consulted for ministries of education, examination boards and international development agencies, and in 2007 joined the Council of Europe as a CEFR expert. She is a recipient of Hornby and IEAA awards.

Nicola Latimer is a Post-Doctoral Research Fellow at the Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire. Her research interests include reading and reading-into-writing, particularly in the academic domain, and the use of eye-tracking to investigate reading processes. She has contributed to several projects including the development of assessment scales for leading exam boards and the development of diagnostic tests for academic reading and writing. Before completing her PhD, she managed a university language support centre for students with English as a second language, and has wide-ranging teaching experience, from entry-level community classes through to teaching academic English at university.

Fumiyo Nakatsuhara is a Reader in Language Assessment at the Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire. She has a PhD in Language Testing and a MA in Applied Linguistics from the University of Essex. Her main research interests lie in the nature of co-constructed interaction in various speaking test formats (e.g. interview, paired and group formats), the impact of test-taker characteristics on test performance, task design, rating scale development, and the relationship between listening and speaking skills. She has carried out a number of international testing projects, working with ministries, universities and examination boards. Her publications include *The Discourse of the IELTS Speaking Test: Interactional Design and Practice* (with Paul Seedhouse, Cambridge Assessment English/Cambridge University Press, English Profile Studies volume 7, 2018) and *The Co-construction of Conversation in Group Oral Tests* (Peter Lang, 2013). Her work also appears in journals such as *Language Testing*, *Language Assessment Quarterly*, *Modern Language Journal* and *System*.

Barry O'Sullivan is the Head of Assessment Research & Development at the British Council where he was responsible for the design and development of the Aptis test service. He has undertaken research across many areas on language testing and assessment and its history and has worked on the development and refinement of the socio-cognitive model of test development and validation since 2000. He is particularly interested in the communication of test validation and in test localisation. He has presented his work at many conferences around the world, while over 100 of his publications have appeared in a range of international journals, books and technical reports. He has worked on many test development and validation projects over the past 25 years and advises ministries and institutions on assessment policy and practice. He was the founding president of the UK Association of Language Testing and Assessment (UKALTA) and holds honorary and visiting chairs at a number of universities globally. In 2016 he was awarded fellowship of

the Academy of Social Sciences in the UK, and was elected to Fellowship of the Asian Association for Language Assessment in 2017. He was awarded an OBE for his contribution to English language testing in 2019.

Jon Roberts worked as a teacher and teacher trainer in Libya, London and Mexico City before taking his MA at the University of Reading in 1977. He was lecturer at the Centre for Applied Language Studies (CALS), University of Reading from then until 2003, when he also completed his doctorate. He is author of *Language Teacher Education* (Arnold, 1998), co-author with Cyril J Weir of *Evaluation in ELT* (Blackwell, 1994), and (with Don Porter) *Authentic listening activities* (*English Language Teaching Journal*, 1981), which is still widely cited.

Lynda Taylor is Visiting Professor at the Centre for Research in English Language Learning and Assessment (CRELLA) at the University of Bedfordshire. She holds a PhD and MPhil in Language Testing from the University of Cambridge and she has worked for many years in the field of language testing and assessment, particularly with IELTS and the full range of Cambridge English Qualifications. Her particular research interests include speaking and writing assessment, test takers with special needs, language assessment literacy, and the role of qualitative methodologies in language testing and assessment. For several years she was Assistant Research Director with Cambridge Assessment English and she has advised on test development and validation projects around the world. She has given presentations and workshops internationally, published extensively in academic journals and authored or edited many of the volumes in the Studies in Language Testing (SiLT) series, co-published by Cambridge Assessment English and Cambridge University Press.

Eddie Williams worked with the British Council in Cyprus and Malta after graduating from Jesus College Oxford, and the Institute of Education, London. He then worked for several years at the Centre for Applied Language Studies (CALS), University of Reading, before moving on to become a Professor at the Department of English Language and Linguistics in the University of Bangor in North Wales. Second language reading, and language in human and economic development (particularly in Africa), were his principal research interests, and he has published widely in these fields. He has been involved in consultancies, workshops and conferences in North and South America, the Middle East and Asia, but above all in Africa, funded by the Department for International Development (DfID), United Nations Educational, Scientific and Cultural Organization (UNESCO), United States Agency for International Development (USAID), European Union (EU) and the British Council, among others. He was an active member of the British

Association of Applied Linguistics (BAAL), serving on various committees and also as Executive Secretary. Now retired, sailing, gardening and grandchildren are his main preoccupations.

Jessica R W Wu holds a PhD in Language Testing. She is currently the R&D Program Director at the Language Training and Testing Center (LTTC), a non-profit educational foundation in Taiwan. She also serves as an advisor to the government on the development and administration of L1 tests. She has published numerous articles and book chapters in the field of language testing and has presented her work at conferences around the world. Most recently she has co-edited and contributed to *English Language Proficiency Testing in Asia: A New Paradigm Bridging Global and Local Contexts* (Routledge, 2019). She is the immediate past president of the Asian Association for Language Assessment (AALA).

Guoxing Yu is Professor of Language Assessment and Director of Research Centre for Educational Assessment and Evaluation at University of Bristol. He is an Executive Editor of *Assessment in Education* (since 2010), and serves on the editorial boards of *Language Testing*, *Language Assessment Quarterly*, *Assessing Writing*, and *Language Testing in Asia*. He is the co-editor of the book series *Language Teachers' Pedagogical Knowledge* (Foreign Language Teaching and Research Press, China). He has published in *Applied Linguistics*, *Applied Linguistics Review*, *Assessing Writing*, *Assessment in Education*, *Educational Research*, *Language Assessment Quarterly* and *Language Testing*. He has directed several research projects on IELTS and the Internet-based Test of English as a Foreign Language (TOEFL iBT). His four IELTS reports are available to download from www.ielts.org, and currently he is working on his fifth research report on IELTS. He earned his PhD in 2005 from Bristol; his dissertation, supervised by Professor Pauline Rea-Dickins, was awarded the Jacqueline A. Ross TOEFL Dissertation Award by Educational Testing Service, USA (2008).

1 A paradigm shift in language testing: determining communicative needs

Vivien Berry
Formerly of British Council

This chapter serves as an introduction to the development of empirical methods of analysis for determining communicative needs. It examines attitudes and beliefs to designing and validating test items in each of the different eras of language testing from the 1970s through to the present day. Topics highlighted in the chapter include:

- the shift from *a posteriori* analysis to *a priori* task design
- the move away from discrete-point, general English proficiency tests to tasks in English for Specific Purposes (ESP) tests
- the shift from tests for different academic disciplines to single tests of overall communicative competence
- the relevance of Weir's early PhD work to studies in the new millennium.

Introduction

Language testing in the early 1970s was firmly rooted in what Spolsky (1977) termed the psychometric-structuralist period. In this period, the focus was on designing test items that could be statistically analysed *a posteriori*. They included item types that were intended to discretely test one language point at a time and were often presented in a multiple-choice or yes–no format. Working at the time within the psycho-structuralist paradigm Davies (1965:52) argued that a language test developer 'starts off from the theory that language can (or should) be analysed into linguistic parts, into language levels'. The test he subsequently developed, the *English Language Test Battery* (often referred to as the *Davies Test*) consequently focused for the most part on lower-order language skills at the decoding level rather than on the higher-order skills of meaning construction and discourse representation.

In the latter part of the 1970s and the early 1980s there was a move away from a simple *a posteriori* approach to test analysis as a way of establishing test reliability and validity through finding out what a learner knew about the

language. Researchers became more concerned with understanding what a learner could do with language in a specific setting, what Spolsky (1977) called the psycholinguistic-sociolinguistic period. Morrow (1979:145) even went so far as to state: 'knowledge of the elements of a language in fact counts for nothing unless the user is able to combine them in new and appropriate ways to meet the linguistic demands of the situation in which he wishes to use the language'. This was the forerunner of communicative language testing, pronounced by Moller (1981) to be the sociolinguistic-communicative paradigm.

One of the main aims of language test designers working within the communicative language testing paradigm was to move away from *a posteriori* test analysis in favour of an *a priori* task development approach which focused on how language is used in real-life situations. As Carroll (1982:1) put it: 'The communicative approach stands or falls by the degree of real-life, or at least life-like, communication that is achieved'. This also involved a move away from the discrete-point structuralist approach to designing general English proficiency tests, into the realms of English for Specific Purposes (ESP), within which academic study can be considered as one of the 'purposes', alongside Medicine, Engineering and other occupational purposes.

Before such tests could be designed, however, there was a need to understand just exactly what language is required in order to communicate effectively in specified linguistic settings. A concerted effort of applied linguists and language testers was therefore required to determine what the actual communicative needs of learners were in specifically identified contexts (cf. *inter alia*, Allwright and Allwright 1977, Candlin 1977, Candlin, Kirkwood and Moore 1978, Carroll 1978, Hawkey 1982, Jones and Roe 1975, Morrow 1977, Munby 1978). Munby's (1978) *Communicative Needs Processor* (based on his 1977 PhD thesis) was particularly influential as it attempted to describe in meticulous, if somewhat unwieldy, detail what overseas students should be able to do in English in specific occupational and academic settings.

Developing a new methodology for communicative needs analysis

Munby's work, especially 'his elaborate mechanism for developing a communicative needs profile' (McNamara 1996:36), has been criticised, notably by Alderson and Hughes (Eds) (1981), Davies (1981), Hawkey (1979) and Skehan (1984), for being impractical and theoretically implausible, as it is essentially an 'armchair' (Alderson 1988:220, Weir 1983a:140) categorisation of needs. In other words it is a cerebral categorisation (Berry 2007:19) or a categorisation made by just sitting and thinking about the problem as opposed to researching it. Nevertheless, his attempt to specify dimensions of performance through his taxonomy of enabling skills was enormously influential in language testing and was used by his colleague at the British Council,

Brendan Carroll, in the development of the successor to the *Davies Test*, namely the British Council's English for academic purposes test, the English Language Testing Service (ELTS) test, first introduced in 1980 (Carroll 1978, 1980).

In addition to the British Council, other examination boards were also developing tests to assess the English language ability of overseas students wanting to study at universities and other higher education establishments in the United Kingdom. An example is the Joint Matriculation Board (JMB) of the Universities of Manchester, Liverpool, Leeds, Sheffield and Birmingham (McEldowney 1976), amongst others. Between 1976 and 1978 the Associated Examining Board (AEB), based in Aldershot, Hampshire, received requests from a number of its teaching centres to develop a test that would provide tertiary institutions with a comprehensive picture of the English language ability of overseas students for whom English was not their first language. In order for them to accede to this request, Cyril Weir was appointed as a research assistant with responsibility for the research and development of a test intended to assess overseas students' readiness to study in English, the AEB's Test in English for Academic Purposes (TEAP).

Weir (1983a) acknowledges the influence of Hawkey (1982), Kelly (1978), Morrow (1977, 1979), and Munby (1978) on the research and development he carried out for the TEAP. As Weir (1983a:112) states:

> We drew upon their research in the construction of a framework of categories for the description of communicative test events: general descriptive parameters, dynamic communicative characteristics and task dimensions of target language behaviour. By applying these categories at the *a priori* test task validation stage we hoped to avoid some of the problems which had arisen in some earlier efforts at communicative testing where no attempt had been made to produce explicit specifications of the candidates' projected language needs in the target situation before test task construction took place . . . we would argue that this approach enabled us to come closer to matching test tasks with appropriate activities in the target behaviour than would be possible using non-empirical approaches.

In order to pursue the communicative paradigm, Weir decided that tasks which would be developed for the test should, as far as possible, relate to the actual tasks, activities, conditions and contexts under which they are normally performed in a tertiary setting. Weir goes on to say (Weir 1983a:112–113):

> The concern was thus with content validity at the *a priori* stage as it no longer seemed sufficient to rely solely on more quantitative *post hoc* validation procedures to establish what it was that we had tested . . . Unless a communicative testing system was initially matched against such a

> framework, it was difficult to see how we could ever get near to describing accurately the construct that we were attempting to measure. The more fully that we could describe the construct through our concern with content validity at the *a priori* stage, the more meaningful were the validation procedures that could subsequently be applied to the results of the test(s).

An empirical needs analysis approach to designing a framework for the realisation of test tasks for the TEAP

An initial provisional framework of descriptive categories of communicative test events was developed, consisting of three phases (see Table 1).

Phase 1 drew on Munby's parameters, developed originally as part of his model for syllabus definition but used by Weir as a checklist against which the appropriacy of performance-based test tasks can be evaluated. Although, as Davies points out, Munby's model is not necessarily 'a blueprint which can be automatically applied' (Davies 1981:332), it could be considered as 'a checklist of things to take into account in determining language communication needs' (Davies 1981:333). Phase 2 owes much to the work of Kelly (1978), Morrow (1977,1979) and Hawkey (1982) in attempting to define how second language learners function in real life in order to make the linguistic activity in the test tasks as appropriate as possible leading to what Widdowson (1978:80) called 'authentic' language use. Phase 3 describes the dimensions of particular events and is essentially derived from Hawkey (1982:166).

Interestingly, Weir states (1983a:121) that he regarded 'Phase 1 as being the most important' and that the data they obtained from Phases 2 and 3 'played a less important role at the test realisation stage'.

Table 1 Framework of categories for the description of communicative test events (from Weir 1983a:114)

Phase 1 **General descriptive parameters of communication**	**Phase 2** **Dynamic communicative characteristics**	**Phase 3** **Task dimensions**
Activities	Realistic content	Size of text
Setting	Relevant information gap	Grammatical complexity and range of cohesion devices required
Interaction	Intersubjectivity	Functional range
Instrumentality	Scope for development of activity by participants	Referential range
Dialect	Allowance for self-monitoring by participants	

Table 1 (continued)

Phase 1 General descriptive parameters of communication	Phase 2 Dynamic communicative characteristics	Phase 3 Task dimensions
Enabling skills	Processing of appropriately sized input Normal time constraints operative	

Once the framework for conducting the research was established, Weir then undertook a systematic series of observations during which he recorded the communication activities the learners were involved in across a variety of disciplines (sciences, engineering, arts, social, administrative and business studies), educational levels (A level, undergraduate and postgraduate) and institutions (universities, polytechnics and A level centres). A separate observation checklist was completed for each lecture, tutorial or practical class in each academic course observed and occurrences were noted on a 4-point scale according to non-occurrence, low, medium and high occurrences. A total of 221 hours of lectures, seminars/tutorials and practical classes were observed. A specimen copy of the observation checklist can be found in Weir (1983a:672–688). The categories dealt with in the observation checklist are listed below (adapted from Weir 1983a:126).

1. **Purpose(s) of study**
2. **Events and activities**
   - 2.1 Lectures
   - 2.2 Seminars/Tutorials
   - 2.3 Practical classes
   - 2.4 Written work
3. **Setting**
   - 3.1 Physical setting: Spatial
   - 3.2 Physical setting: Temporal
   - 3.3 Psychosocial setting (i.e. operating in the quiet of a library or seminar room or the noisier atmosphere of a workshop)
4. **Interactions (i.e. student–student or student–teacher centred, etc.)**
   - 4.1 Position
   - 4.2 Role set
   - 4.3 Role set identity
   - 4.4 Inventory of social relationships

5. **Instrumentality (i.e. spoken productive, spoken to be written, face-to-face, etc.)**
   - 5.1 Medium
   - 5.2 Mode
   - 5.3 Channel
   - 5.4 Non-verbal medium
6. **Target level (i.e. complexity, speed, repetition, hesitation, etc.)**
   - 6.1 Dimensions
   - 6.2 Tolerance conditions
7. **Communicative key (i.e. the attitudinal tone in which the event is carried out)**

Whenever possible, an opportunistic sample of teachers and students was interviewed after the observations. The purpose of the interviews was to gain further information which had been impossible to get during the actual observations, to check that the data gathered in the observations was generalisable to the entire course, to establish the main English language problems experienced by both overseas and British students on the course and, finally, to examine the efficiency of the questions developed for the staff and student pilot questionnaires. These empirically based techniques culminated in the development of the major data collection instrument, a questionnaire survey.

Once all the data gathered from the observations and interviews was analysed, the pilot staff and student questionnaires were refined and final questionnaires were produced. A total of 5,947 final questionnaires were then sent out to a variety of institutions and distributed to staff, overseas students and British students. One thousand nine hundred and forty completed questionnaires were returned (950 overseas students, 430 British students, 560 staff) from 43 postgraduate courses, 61 undergraduate courses and 39 A level centres. This resulted in a basic core of empirical evidence which provided a specification of the actual academic contexts in which the students would have to operate, and also established through the questionnaire the academic activities overseas students had most difficulty with compared to the British students.

Five parameters were derived from the data and used to inform the test task construction phase which followed. The parameters (from Weir 1983a:144) were:

a. purpose(s) (of the participant(s) in the event and of the event itself – i.e. lectures, seminars/tutorials, practical classes)
b. activities (sub-tasks involved in achieving the purpose(s) – i.e. listening comprehension, note-taking, reading comprehension, writing and speaking activities in the academic context)

c. setting (physical and psychosocial – i.e. physical environment, spread of hours per week in various learning situations and psychosocial environments they operated in)
d. interaction (role set and social relationships of participants)
e. instrumentality (medium, mode, channel of communication of the event).

Although Weir included a section on speaking in the final questionnaires completed by A level, undergraduate/postgraduate students and staff, a decision had been taken by the AEB at an early stage in the project that the TEAP would not initially have an oral component (see Weir (1983a:272) for an explanation of the AEB decision). Weir (1983a, 1983b, 1988) therefore focuses only on the development of the TEAP listening, reading and writing tasks.

A trial test was developed and administered, results were analysed, and a final TEAP test consisting of discrete, integrated and integrative tasks was designed to be administered by the AEB. When it became operational, the TEAP was rebranded as the Test in English for Educational Purposes (TEEP). For full details of the later development of the TEEP's semi-direct speaking test, see James (1988:111–133).

The TEEP test was designed in the early 1980s at around the same time as the ELTS test was first rolled out (1980). Despite the TEEP test having been thoroughly researched and developed specifically to provide a measure of overseas students' ability to follow a course at an English-medium university, described by Alderson (1988:223) as 'this monumental work', it could not compete with ELTS (and the revised version, IELTS). After the AEB sold the TEEP test to the University of Reading in the 1980s, it was hardly ever used outside that university and only between 1,000 and 1,500 TEEP tests are delivered annually. This compares with over 3.5 million IELTS test taken in 2018 in over 1,200 centres in 140 countries in the world. As Weir and O'Sullivan commented recently (2017:161): 'Unfortunately for TEEP there was no organisation like the British Council or IDP [International Development Program of Australian Universities] to promote the exam on the global stage or deliver it effectively overseas through an efficient and widespread infrastructure'.

Nevertheless, although TEEP was not destined to become a major global academic language test, Weir's work on developing the test tasks for it would remain influential in language testing to this day. In the rest of this chapter, we will look at some of the projects that were influenced by Weir's methods of needs analysis.

Using empirical needs analysis in the redevelopment of the Occupational English Test (OET)

Until 1987 the Occupational English Test (OET) was a test of general proficiency taken by immigrants and refugees to Australia who had previously qualified as health professionals overseas (mostly doctors but also nurses, dentists, physiotherapists, occupational therapists, speech pathologists, veterinary surgeons, etc.) in non-English medium contexts. The original test had attracted criticism from test takers and test users alike who questioned its content validity. As a result of this criticism, several consultancies were set up to reform the test. The first consultancy was carried out by a team from Lancaster University, which recommended designing a new test to 'assess the ability of candidates to communicate effectively in the workplace' (Alderson, Candlin, Clapham, Martin and Weir 1986:3). This recommendation in effect suggested that the OET proficiency test should be redesigned as a performance test. An attempt was made to operationalise the recommendation, after which a further series of reports were published (McNamara 1987, 1988, 1989).

In work sample tests, content selection is crucial in respect to content validity. Davies (1977:62) contends that establishing content validity involves the following: 'An assessment must be made of just what the learners whose proficiency is to be tested need to do with the language, what varieties they must employ and in what situations they must use those varieties.'

In order to conduct a job analysis for reforming the OET, McNamara, following Weir's (1983a, 1983b) work on the TEEP (originally the TEAP), used a number of procedures including interviews with those involved in the professional training of both local and overseas health professionals to draw up a tentative list of work-related communicative tasks. McNamara drew on Weir's (1988) practical approach of asking his informants which communicative tasks were most frequent, which were most complex or difficult and which were most important. In this case his informants were migrant doctors who had completed registration as medical professionals in Australia and were working in clinical settings, as these are by far the largest group of health professionals required to take the OET. The data acquired from the interviews was used to develop a questionnaire which was then administered to overseas medical graduates. Unlike Weir's results which informed the design of the TEAP/TEEP, McNamara's results showed that the 10 most frequent tasks in the workplace were oral, the most frequent being face-to-face communication with patients and their families. This result enabled designers of the reformed OET to insist on a live speaking subtest as well as informing the broad content parameters for the speaking and other skills subtests.

Following analysis of the questionnaire responses, time was then spent observing actual workplace communication in all of the different professions

which the test catered for. This was done in an attempt to establish commonalities between the professions as a basis for test task design. The following commonalities were perceived across all professions (from McNamara 1996:104):

1. Assessment of the patient ("subjective assessment") including history
2. Physical examination
3. Explanation to the patient of diagnosis and prognosis and course of treatment
4. Treatment
5. Patient/client/relative education and counselling

These commonalities allowed for the writing of materials which were then trialled, analysed and revised. Raters were recruited and trained. Analysis of data received from the trials was conducted, and test materials and specifications were revised. Minimum acceptable performance standards were determined and implementation and monitoring of the revised test was undertaken. Since many of these procedures are common to the development of any new test, they will not be further elaborated here. For further information about the redevelopment of the OET, see McNamara (1990).

Target-centred needs analysis such as that of Weir with TEAP/TEEP and McNamara with the OET was modelled on those academic study skills or job-related activities that second language speakers would have to cope with in their studies or workplace setting. This 'real-life' approach was at the time derided by the more psychometrically inclined language testers in the United States. For example, Bachman, discussing the challenges that communicative approaches to language testing present, states (1990:299):

> These challenges also present an opportunity for us to move the field of language testing forward, to reforge the symbiotic relationship between applied linguistic theory and the tools of psychometrics and statistics that was characteristic of the "psychometric-structuralist" trend in language testing (Spolsky 1981), and that has, in my opinion, been largely lost sight of in the current "integrative-sociolinguistic" trend.

However, as Weir and McNamara have repeatedly shown, with extreme care in the design, development and ongoing monitoring of actual test performance, including the use of new measurement models such as Many-facet Rasch measurement (MRFM) (Linacre 1989), situational and interactional authenticity can be achieved. This is especially true of the work of McNamara and Lumley (1993) and McNamara (1996), who offered the first comprehensive presentation of Rasch measurement, an approach that enables investigation of aspects of performance settings such as rater and task characteristics and is now commonly used in every language testing context. The irony of

this was not lost on Weir and O'Sullivan (2017:77) who comment: 'Ironically the pursuit of such situational and interactional authenticity has now become the approach of choice of leading American testers as in the new TOEFL iBT (www.ets.org/toefl/ibt/about).'

Needs analysis and the redevelopment of the ELTS test, leading to the birth of IELTS

As mentioned earlier, the TEEP was developed in the early 1980s at about the same time as the ELTS test was introduced. The influence of needs analysis and communicative language demands in study or work contexts meant that subtests of writing and speaking ability were included in the new tests. However, unlike the TEEP, which had a semi-direct speaking component (see James 1988), the ELTS test included speaking and writing components which were assessed directly (writing two essays from prompts and a three-part one-to-one oral interview).

Using a direct approach to assessing productive skills presents test developers with considerable theoretical and practical challenges, not the least of which concerns the recruitment of qualified teachers who then have to be trained to mark the essays and to conduct and rate the oral interviews. Another challenge was that the original ELTS test was very difficult to administer because of its design as an ESP test with six different academic domains (Life Sciences, Social Studies, Physical Sciences, Technology, Medicine, General Academic). From the beginning there were numerous complaints about its unwieldiness and the time needed to administer it (a total of 175 minutes to complete all five sections of the test). It was also heavily criticised on numerous grounds by several experienced language testing experts at the inaugural meeting of the Language Testing Forum in 1980 (Alderson and Hughes (Eds) 1981). As a result of all this criticism, a team from the University of Edinburgh proposed to conduct a validation project to provide information for test users with information on the test's fitness for purpose.

In the early 1980s it was unusual to instigate a validation study so soon after a new test had been operationalised, but ELTS offered an innovative approach to communicative language testing and it is quite probable that the test developers wanted to share their evidence on how well it was working in comparison with more traditional test types. So, the validation proposal was approved and the team from Edinburgh was invited to conduct the ELTS Validation Study (ELTSVAL 1981–86).

The specific aims of ELTSVAL, taken from Westaway, Alderson and Clapham (1990:241), were:

1. To examine the predictive validity of ELTS in relation to students' success in their academic studies.

2. To examine the face, content and construct validity of ELTS.
3. To examine the concurrent validity between ELTS, English Proficiency Test Battery (EPTB) and University of Edinburgh English Language Battery (ELBA).
4. To assess the extent to which proficiency in English affects success in academic studies.
5. To investigate the internal reliability and retest reliability of ELTS.

Discussing the results of the ELTSVAL study, Criper and Davies (1988:114), maintain that:

> . . . it is a satisfactory test of English proficiency because of its reliability and certain claims on validity. Its face validity is high but its content validity is less so. In terms of construct validity our evidence from the predictive and concurrent studies suggests specialists do ideally require different subtests or combinations of subtests but that the model presented in the present ELTS tests of specialist modules is not effective . . . Similarly, in practical terms, our concurrent and predictive studies indicate that a shorter and more easily administered test would be equally effective.

Following the report on the ELTSVAL project, the ELTS Revision Project (1986–89) was instigated under the direction of Charles Alderson and a team from Lancaster University. Having been critical of Carroll's and Weir's needs analyses for the original ELTS and TEEP tests (cf. Alderson 1988, Clapham 1981), the Lancaster team attempted to devise a different methodology to address the aims of the Revision Project. However, despite their earlier criticisms of Carroll's Munby-like specifications for ELTS and Weir's use of Munby's needs analysis for TEEP, Alderson and Clapham (1992:163) admit: 'Since several analyses have been carried out into the language needs of tertiary-level students (in particular Weir 1983[a]), we used these for the test specifications and tests'.

A full description of the processes and procedures involved in the ELTS revision is contained in Alderson and Clapham (1997). It will suffice here to say that the major revisions which emerged from the 1986–89 Revision Project came about as a result of the addition of a third partner, the International Development Program of Australian Universities (IDP Australia) who had joined the British Council and UCLES as co-owners of the test. This resulted in the change in name from the English Language Testing Service (ELTS) test to the International English Language Testing System (IELTS) test. There was also a reduction in time from 180 minutes to a new total time of 110 minutes to cater for the criticisms relating to practicality that had been levelled at ELTS. In addition, there was a reduction in the number of academic modules from six to three – Physical Sciences

and Technology (PST), Life and Medical Sciences (LMS) and Business Studies and Social Sciences (BSS). These were themselves reduced to a single academic module in the 1995 IELTS revision thus concurring with Weir's (1983a:549–550) conclusion:

> In our investigations of the language events and activities overseas students have to deal with in British academic environments and the difficulties they encounter therein, we discovered much that was common between students of different disciplines and at different levels . . . we were unable to produce any conclusive evidence that students were disadvantaged by taking tests in which they had to deal with texts other than those from their own subject area. The case for a variety of ESP tests therefore remains unproven.

For a fuller understanding of the historical development of ELTS/IELTS, see also Clapham (1996), Davies (2008), Taylor and Falvey (Eds) (2007) and Taylor and Weir (Eds) (2012).

The continuing relevance and role of empirical needs analysis in the new millennium

So far in this chapter I have outlined a number of projects where the researchers derived their methodology from Weir's empirical needs analysis, developed for his PhD thesis in the early 1980s, almost 40 years ago. In this final section I will look at a project conducted very recently in order to illustrate how Weir's methodology is still relevant today. The earlier projects typically involved needs analysis focusing on the requirements of academic study or professional workplace contexts. The following example, however, concerns a new area relating to language and teacher education, demonstrating the extent of Weir's influence in a variety of domains.

One of the recent trends in language assessment has been a focus on language assessment literacy, especially with regard to the language assessment literacy of language teachers. For an overview of some of the issues, see the collection of articles in the 2013 Special Issue of *Language Testing*, issue 30 (3) edited by Ofra Inbar-Lowrie. See also Berry, O'Sullivan, Schmitt and Taylor (2014), Berry, Sheehan and Munro (2017a, 2017b, 2019), Crusan, Plakans and Gebril (2016), Fulcher (2012), Malone (2011), *inter alia*.

Effective assessment can support and promote learning, and therefore a teacher's ability to engage with a range of teaching, learning and assessment practices is essential. However, concerns have been expressed about the level and quality of teacher training in assessment (Crusan et al 2016, Fulcher 2012, Vogt and Tsagari 2014). In general education, the term assessment literacy has been used to describe the knowledge teachers should have about

assessment. The term has been adapted and adopted by experts in language assessment, with Malone (2011) proposing the following definition of language assessment literacy: 'Assessment literacy is an understanding of the measurement basics related directly to classroom learning; language assessment literacy extends this definition to issues specific to language classrooms'.

The survey has been the most commonly used research method when investigating teachers' knowledge of assessment (Crusan et al 2016, Fulcher 2012). These surveys are generally created by expert researchers in assessment, often leading to an emphasis on possible gaps in teacher knowledge. In many studies, a range of assessment-related topics is presented to teachers with questions that ask them to state their current level of knowledge about the topic and their interest in learning more about it. The results of many of these surveys suggest that teachers lack knowledge and need extensive training to raise their level of understanding of these areas of assessment. Survey studies, however, have limitations, not the least of which is that they are still often constructed in the manner which Weir (1983a) and Alderson (1988) have described as 'armchair' categorisations.

Additional limitations include the fact that respondents to surveys, especially online surveys, are probably self-selected from among those who are interested in the topic in the first place. Second, teachers' responses may reflect what they think they should say, rather than what they actually believe. A corollary to this is that training needs may be exaggerated in the belief that it would appear unprofessional to state that they had no interest in the topic. Also, affirmative answers may be given out of curiosity rather than genuine interest or need to know. And following data collection, interpretation of responses may rely too heavily on quantitative analysis at the expense of individual differences.

Berry et al's 2016–19 study sought to minimise the limitations of survey studies by conducting in-depth initial interviews, classroom observations with follow-up interviews and finally a series of focus group discussions, in order to gain empirically derived insights into what teachers actually do in terms of assessment in classrooms. The aims of the study were therefore twofold (from Berry et al 2019):

1. To gain a greater understanding of teachers' knowledge of assessment through actual observation of classroom assessment practices and through focus group discussions.
2. To use the knowledge gained from the observations and discussions to develop training materials which meet teachers' actual stated needs.

Participants in the study were teachers who were based in Europe at the time of the project but many of them talked about work and training experiences from beyond Europe in both state education and private language schools. A total of 54 teachers participated in the study, 28 of them female and 26

male, with ages ranging from 25 to 60 years. The teachers were chosen, from amongst those who had volunteered to participate in the study, to reflect a range of routes into teaching and to include teachers working in different contexts in order to ensure that our findings were as generalisable as possible to a larger population of teachers.

The study comprised three phases. Phase 1 consisted of a series of interviews with three experienced, international EFL teachers, conducted in the School of Education at a British university. The interviews drew on Davies' (2008:335–341) components of assessment literacy, *Skills* + *Knowledge* + *Principles*, which he defined as follows:

> Skills provide the training in necessary and appropriate methodology, including item writing, statistics, test analysis and increasingly software programmes for test delivery, analysis and reportage. Knowledge offers relevant background in measurement and language description, as well as in context setting, and may involve an examination of different models of language learning, of language teaching and of language testing such as communicative language testing, performance testing and nowadays, socio-cultural theory. Principles concern the proper use of language tests, their fairness and impact, including questions of ethics and professionalism . . .

A more succinct discussion of skills, knowledge and principles is offered by Taylor (2013) in her commentary paper on some of the themes originally raised during the 2011 Language Testing Research Colloquium (LTRC) symposium on assessment literacy. Taylor (2013:410) summarised the eight components she identified as:

- knowledge of theory
- technical skills
- principles and concepts
- language pedagogy
- sociocultural values
- local practices
- personal beliefs/attitudes
- scores and decision making.

In Phase 1 of the Berry et al study, teachers were invited to estimate their understanding of each of Taylor's components and were also asked about their own experiences of assessment and how they had developed their assessment practices. They discussed their initial teacher training and other training opportunities they had had.

In Phase 2, observations were conducted in the International Study Centre of the same university, which focused on teachers' actual assessment practices

in the classroom. Using an observation schedule inspired by Colby-Kelly and Turner's (2007) study of assessment for learning practices, a checklist of 16 assessment practices was developed, and every 3 minutes during the observations, checks were noted of which of the practices were being observed and notes were written about them. Post-observation interviews were subsequently conducted with the three teachers, in which they were asked to reflect on their observed classroom practice and discuss why they had used particular assessment techniques in class. The observed teachers were not the same as the ones interviewed in Phase 1.

Finally, in Phase 3, focus group discussions were held with 48 experienced teachers working at teaching centres attached to a major international organisation in Madrid and Paris. These teachers taught a variety of different English language classes across a range of students, plus special-purpose classes for commercial organisations. Although as previously stated the teachers were all based in Europe at the time, they had a huge amount of experience working in all parts of the world. The main purpose of this stage was therefore to confirm that the comments from Phases 1 and 2 were typical of a much broader range of English language teachers.

Data analysis drew on the components of assessment literacy detailed above. Three key findings emerged from the analysis relating to teachers' previous training in assessment, attitudes to language testing and assessment in its broader sense, and the types of training materials they would like. In discussion, teachers acknowledged their lack of training, but suggested that the divide between teaching and assessment begins in pre-service training when teaching is prioritised and assessment is not considered to be important. In terms of their attitudes to language testing and assessment, participants referred to 'testing' rather than 'assessment' although the researchers always used the term 'assessment'.

The lack of engagement with assessment may be a consequence of the limited role some teachers play in the development and creation of assessments. This would seem to provide support for the notion that teachers feel assessment is a top-down imposition (Crusan et al 2016). In addition, there is some evidence to suggest that testing is only acceptable if it can be used to support or improve teaching in some way. This is a further demonstration of the gap between teaching and assessment as teaching is being privileged. Experience, rather than training, seems to play a pivotal role in the development of assessment practices. This experience develops with time spent in the classroom.

Regarding the training materials requested, most of the teachers who participated in the study expressed their training needs in terms of requests for activities and not in terms of theory or principles, thus confirming Davies' (2008) claim that there is little demand for theory among teachers. Teachers mainly requested training materials related to skills, tasks and criteria,

including video examples and particularly more practical elements that were situation-based. This is in direct contrast to the types of training materials language assessment experts often suggest teachers need. It may also be that there is a disconnect between teachers' interests and beliefs and those of language assessment professionals and researchers. Berry et al's (2017a, 2017b, 2019) findings suggest that the gap between teachers and those who research and write about language testing is considerable.

Conclusions

We saw in the first part of this chapter how Weir's work as a research assistant with the AEB investigating the educational needs of overseas students wishing to study in the UK led to the development of the Test in English for Academic Purposes (TEAP). This test, later rebranded by AEB as the Test in English for Educational Purposes (TEEP), consisted of discrete, integrated and integrative tasks. Weir's insistence in his early work on the importance of prioritising empirical needs analysis to underpin test design, development and validation, remains the model for developing tests of English for academic purposes today.

In addition to providing a model for creating tests for educational purposes in the UK, Weir's work also directly influenced researchers such as McNamara, who acknowledges that he drew on Weir's practical approach of interviewing informants and conducting follow-up observations when redesigning the Occupational English Test (OET), a workplace test for overseas trained professionals wishing to practise their professions in Australia. Target-centred needs analysis such as that of Weir with the TEAP/TEEP and McNamara with the OET was modelled on the academic skills or job-related activities that non-English first language speakers would have to cope with in their studies or workplace setting. Although complicated and expensive to develop, both Weir and McNamara have repeatedly shown that with extreme care in the design, development and ongoing monitoring of actual test performance, situational and interactional authenticity can be achieved in tests of English for specific purposes.

In the third section of the chapter, what can probably be considered as one of Weir's major contributions to language assessment was briefly described. This is, of course, his involvement in the validation and revision of the ELTS test, with Alderson and Clapham (1992), drawing on what Alderson (1988:223) had earlier termed 'this monumental work', to develop the specifications for the revised ELTS test, which later became the IELTS test. Interestingly, the ELTS test originally had six domain-specific modules. When IELTS was created in 1989, these six modules were reduced to three domain-specific modules plus a general training module. The revision of IELTS in 1995 then saw the reduction of the three domain-specific

modules to a single academic module, plus one general training module. This agrees with Weir's (1983a) finding that there was no evidence to suggest that students were disadvantaged through having to deal with a topic outside their immediate sphere of interest. So, from starting out in the early 1980s with a limited audience in higher education institutions in the UK, Weir's influence towards the end of the second decade of the 21st century now extends throughout the world. According to the latest figures available, 3.5 million international IELTS tests were taken in 140 countries worldwide in 2018.

The final piece of research described in this chapter also attests to the continuing impact of Weir's legacy. In a study of teachers' assessment literacy, Berry et al, following Weir's, and later, McNamara's methodology, conducted initial interviews with teachers and then developed checklists for observations of performances in the workplace. In this instance, the workplace was also an educational setting (classroom) as the observations were of teachers instructing their students. Berry et al's findings suggest that what teachers think they need in terms of assessment literacy and what language assessment experts say they need are often quite far apart, confirming that surveys based on 'armchair categorisations' of needs bear little relationship to the actual needs of practitioners.

To conclude, in keeping with that which Cyril Weir (1983a) determined so long ago, prioritising empirical needs analysis to underpin test design, development and validation allows for tests to be created that are appropriate for specific populations. As the studies presented in this chapter show, it also allows for the development of materials that are relevant to the needs of individual participants. It is clear, therefore, that empirical needs analysis will be the methodology of choice for language test developers for many years still to come.

References

Alderson, J C (1988) New procedures for validating proficiency tests of ESP? Theory and practice, *Language Testing* 5 (2), 220–232.

Alderson, J C and Clapham, C M (1992) *IELTS Research Reports 2: Examining the ELTS Test: An Account of the First Stage of the ELTS Revision Project*, Cambridge: British Council/UCLES/IDP.

Alderson, J C and Clapham, C M (1997) *IELTS Research Reports 3: Constructing and Trialling the IELTS Test*, Cambridge: British Council/UCLES/IDP.

Alderson, J C and Hughes, A (Eds) (1981) *E.L.T. Documents 111 – Issues in Language Testing*, London: British Council.

Alderson, J C, Candlin, C N, Clapham, C M, Martin, D J and Weir, C J (1986) *Language proficiency testing for migrant professionals: New directions for the Occupational English Test. A report submitted to the Council on Overseas Professional Qualifications*, Lancaster: Institute for English Language Education, University of Lancaster.

Allwright, J and Allwright, R (1977) An approach to the teaching of medical English, in Holden, S (Ed) *English for Specific Purposes,* Oxford: Modern English Publications, 58–62.

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

Berry, V (2007) *Personality Differences and Oral Test Performance*, Frankfurt am Main: Peter Lang.

Berry, V, Sheehan, S and Munro, S (2017a) *Beyond surveys: an approach to understanding effective classroom assessment practices*, paper presented at IATEFL TEASIG/CRELLA seminar, University of Bedfordshire, Luton, 28–29 October 2017.

Berry, V, Sheehan, S and Munro, S (2017b) Exploring teachers' language assessment literacy: A social constructivist approach to understanding effective practices, in ALTE (Ed) *Learning and Assessment: Making the Connections – Proceedings of the ALTE 6th International Conference*, 201–207.

Berry, V, Sheehan, S and Munro, S (2019) What does language assessment literacy mean to teachers?, *ELT Journal* 73 (2), 113–123.

Berry, V, O'Sullivan, B, Schmitt, D and Taylor, L (2014) *Assessment Literacy: Bridging the Gap between Needs and Resources*, panel discussion at the 48th IATEFL Conference, Harrogate, 2–5 April 2014.

Candlin, C N (1977) *Questionnaire on the English language problems faced by overseas students studying in Britain*, unpublished internal mimeo, University of Lancaster, Institute of English Language Education.

Candlin, C N, Kirkwood, J M and Moore, H M (1978) Study skills in English: Theoretical issues and practical problems, in Mackay, R and Mountford, A (Eds) *English for Specific Purposes: A Case Study Approach*, London: Longman, 190–219.

Carroll, B J (1978) *An English Language Testing Service: Specifications*, London: British Council.

Carroll, B J (1980) *Testing Communicative Performance: Principles and Practice*, Oxford: Pergamon Press.

Carroll, B J (1982) Language testing – is there another way?, in Heaton, J B (Ed) *Language Testing*, London: Modern English Publications, 1–10.

Clapham, C M (1981) Reaction to the Carroll paper (1), in Alderson, J C and Hughes, A (Eds) *E.L.T. Documents 111 – Issues in Language Testing*, London: British Council, 111–116.

Clapham, C M (1996) *The Development of IELTS*, Studies in Language Testing volume 4, Cambridge: UCLES/Cambridge University Press.

Colby-Kelly, C and Turner, C (2007) AFL research in the L2 classroom and evidence of usefulness: Taking formative assessment to the next level, *Canadian Modern Language Review* 64 (1), 9–37.

Criper, C and Davies, A (1988) *ELTS Research Report 1(i): ELTS Validation Project Report*, Cambridge: British Council/UCLES.

Crusan, D, Plakans, L and Gebril, A (2016) Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices, *Assessing Writing* 28, 43–56.

Davies, A (1965) *Proficiency in English as a Second Language*, unpublished PhD thesis, University of Birmingham.

Davies, A (1977) The construction of language tests, in Allen, J P B and Davies, A (Eds) *Testing and Experimental Methods. The Edinburgh Course in Applied Linguistics*, Oxford: Oxford University Press, 38–104.

Davies, A (1981) A review of Communicative Syllabus Design, *TESOL Quarterly* 15 (3), 332–344.
Davies, A (2008) Textbook trends in teaching language testing, *Language Testing* 25 (3), 327–347.
Fulcher, G (2012) Assessment literacy for the language classroom, *Language Assessment Quarterly* 9 (2), 113–132.
Hawkey, R (1979) Communicative needs profiles, in Abbot, G and Beaumont, M (Eds) (1997) *The Development of ELT: the Dunford Seminars 1978–1993*, Hemel Hempstead: Prentice Hall Europe, 23–30.
Hawkey, R (1982) *An Investigation of Inter-relationships between Cognitive/ Affective and Social Factors and Language Learning. A Longitudinal Study of 27 Overseas Students Using English in Connection with Their Training in the United Kingdom*, unpublished PhD thesis, University of London.
Inbar- Lawrie, O (Ed) (2013) *Language Testing* 30 (3).
James, G (1988) Development of an oral proficiency component in a test of English for academic purposes, in Hughes, A (Ed) *Testing English for University Study*, ELT Documents 127, London: Modern English Publications/British Council, 111–133.
Jones, K and Roe, P (1975) Designing English for Science and Technology (EST) programmes, in British Council, *English for Academic Study with Special Reference to Science and Technology. Problems and Perspectives*, London: British Council, 1–45.
Kelly, R (1978) *On the construct validation of comprehension tests: An exercise in applied linguistics*, unpublished PhD thesis, University of Queensland.
Linacre, J M (1989) *Many-facet Rasch Measurement*, Chicago: MESA Press.
Malone, M E (2011) Assessment literacy for language educators, *CAL Digest October 2011*, available online: www.cal.org
McEldowney, P L (1976) *Test in English (Overseas)*, *The Position after Ten Years*, Occasional Paper 36, Manchester: Joint Matriculation Board.
McNamara, T F (1987) *Assessing the Language Proficiency of Health Professionals. Recommendations for Reform of the Occupational English Test. A Report Submitted to the Council on Overseas Professional Qualifications*, Melbourne: University of Melbourne.
McNamara, T F (1988) *The Development of an English as a Second Language Speaking Test for Health Professionals. Part One of a Report to the Council on Professional Overseas Qualifications on a Consultancy to Develop the Occupational English Test*, Melbourne: University of Melbourne.
McNamara, T F (1989) *The Development of an English as a Second Language Writing Test for Health Professionals. Part Two of a Report to the Council on Professional Overseas Qualifications on a Consultancy to Develop the Occupational English Test*, Melbourne: University of Melbourne.
McNamara, T F (1990) *Assessing the Second Language Proficiency of Health Professionals*, unpublished PhD thesis, University of Melbourne.
McNamara, T F (1996) *Measuring Second Language Performance*, London: Longman.
McNamara, T F and Lumley, T (1993) *The Effects of Interlocutor and Assessment Mode Variables in Offshore Assessment of Speaking Skills in Occupational Settings*, paper presented at the 15th Annual Language Testing Research Colloquium, Cambridge, August 1993.
Moller, A D (1981) Reaction to the Morrow paper (2), in Alderson, J C and Hughes, A (Eds) *E.L.T. Documents 111 – Issues in Language Testing*, London: British Council, 38–44.

Morrow, K (1977) *Techniques of Evaluation for a Notional Syllabus*, London: The Royal Society of Arts.

Morrow, K (1979) Communicative language testing: Revolution or evolution?, in Brumfit, C J and Johnson, K (Eds) *The Communicative Approach to Language Teaching*, Oxford: Oxford University Press, 143–158.

Munby, J L (1977) *Designing a Processing Model for Specifying Communicative Competence in a Foreign Language: A Study of the Relationship Between Communicative Needs and the English Required for Specific Purposes*, unpublished PhD thesis, University of Essex.

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

Skehan, P (1984) Construct validity, in Hughes, A, Porter, D and Weir, C J (Eds) *ELTS Research Report 1(ii): ELTS Validation Project: Proceedings of a Conference Held to Consider the ELTS Validation Project Report,* Cambridge: UCLES, 26–31.

Spolsky, B (1977) Language Testing: Art or Science?, in Nickel, G (Ed) *Proceedings of the Fourth International Congress of Applied Linguistics*, Stuttgart: Hochschulverlag, 7–28.

Taylor, L (2013) Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections, *Language Testing* 30 (3), 403–512.

Taylor, L and Falvey, P (Eds) (2007) *IELTS Collected Papers 1: Research in Speaking and Writing Assessment*, Studies in Language Testing volume 19, Cambridge: UCLES/Cambridge University Press.

Taylor, L and Weir, C J (Eds) (2012) *IELTS Collected Papers 2: Research in Reading and Listening Assessment*, Studies in Language Testing volume 34, Cambridge: UCLES/Cambridge University Press.

Vogt, K and Tsagari, D (2014) Assessment literacy of foreign language teachers: Findings of a European study, *Language Assessment Quarterly* 11 (4), 374–402.

Weir, C J (1983a) *Identifying the language problems of the overseas students in tertiary education in the United Kingdom*, unpublished PhD thesis, University of London.

Weir, C J (1983b) The Associated Examination Board's Test in English for Academic Purposes: An exercise in content validation, in Hughes, A and Porter, D (Eds) *Current Developments in Language Testing*, London: Academic Press, 147–153.

Weir, C J (1988) The specification, realization and validation of an English language proficiency test, in Hughes, A (Ed) *Testing English for University Study*, London: Modern English Publications/British Council, 45–110.

Weir, C J and O'Sullivan, B (2017) *Assessing English on the Global Stage: The British Council and English Language Testing 1941–2016*, Sheffield: Equinox.

Westaway, G, Alderson, J C and Clapham, C (1990) Directions in testing for specific purposes, in de Jong, J H A L and Stevenson, D K (Eds) *Individualising the Assessment of Language Abilities*, Bristol: Multilingual Matters, 239–256.

Widdowson, H G (1978) *Teaching Language as Communication*, Oxford: Oxford University Press.

Foreword to Chapter 2

Barry O'Sullivan, British Council, July 2019

This chapter is a much-abbreviated version of the 2002 report that saw the first airing of the ideas that were to move from the edge of language testing (I was asked at a Language Testing Forum at the time why we were interested in cognition) to its core. The reason I was keen to include a version of this report was that it shows the genesis of the socio-cognitive framework which has become so instrumental in recent years for many parts of the language testing community. It also helps us to understand that the original frameworks (and later model) built on our discussions in the late 1990s/early 2000s, and was helped hugely by our PhD students and colleagues in the language testing community.

The idea of a socio-cognitive approach to developing and validating tests of productive language first came to me pretty late in the process of writing my PhD thesis. That would put it in late 1999. Though I discussed the idea with my supervisor (Don Porter) and with Cyril at the time, it wasn't until a year or so later that we began to think seriously about it. At the time, my primary interest was in the test taker and the social aspect of language use in the test event – i.e. how a test taker's language use was affected by their interlocutor or audience/reader. Cyril, on the other hand, had become increasingly interested in the cognitive behaviour of test takers – i.e. to what extent did the test task elicit the same processes as an individual might employ in performing such a task in a non-test context?

We had been working independently when we were both asked to speak at an ALTE Conference Day in St Petersburg in April 2002. As we were parked on the tarmac while the technicians worked on our Lufthansa flight's weather radar, we took out a napkin and, armed with a cheap pen, we sketched out our ideas, linking the 'socio' to the 'cognitive' in what was to become the first of the socio-cognitive frameworks. While we both contributed to that initial sketch, it was Cyril's genius to conceive of a practical framework consisting of a number of interacting elements. The reason we were interested in developing a speaking framework was that we were, at that time, working on a report commissioned by Cambridge Assessment English (then UCLES) on research issues in testing speaking. The basic structure (and elements) stayed the same from that early stage, leading to Cyril's iconic 2005 book, *Language Testing and Validation: An Evidence-based Approach.* It helped to inform the test development and validation research agenda at Cambridge

Assessment English and the subsequent series of Studies in Language Testing (SiLT) publications – *Examining Writing*/*Reading*/*Speaking*/*Listening*. As the approach evolved, it also inspired many international research and development projects, including the British Council's International Language Assessment (ILA) placement test and the later Aptis test system. Despite the later changes we made (O'Sullivan and Weir 2011, O'Sullivan 2016) to facilitate the development of the underlying model – the primary focus these days being on three core elements (test taker, test system, scoring system) all located in a specific context of development and use – the basic tenets of the approach remain as when they were first envisaged.

Cyril and I continued to work on the frameworks, him primarily on the cognitive aspects and me on attempting to define the underlying model and to operationalise the concept of consequence. We met regularly to discuss everything from rugby (except when Ireland managed to beat England) to the language testing world, talking about the new books, papers and ideas that were influencing (or not) people's thinking. At the time of his death in 2018, we were working on plans to jointly update the 2005 book, again bringing the different strands of our work together. Sadly, it was not to be.

As you will see here, the ideas may have been put together in a new and exciting way, but they were 'built on the shoulders of giants', as our friend Micheline Chalhoub-Deville informed us on many occasions. Now that the greatest giant of them all is no longer here, it is incumbent on us all to continue to build on the ideas he set out so well.

References

Geranpayeh, A and Taylor, L (Eds) (2013) *Examining Listening: Research and Practice in Assessing Second Language Listening*, Studies in Language Testing volume 35, Cambridge: UCLES/Cambridge University Press.

Khalifa, H and Weir, C J (2009) *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing volume 29, Cambridge: UCLES/Cambridge University Press.

O'Sullivan, B (2016) Validity: What is it and who is it for?, in Leung, Y (Ed) *Epoch Making in English Teaching and Learning: Evolution, Innovation, and Revolution*, Taipei: Crane Publishing Company Ltd, 157–175.

O'Sullivan, B and Weir, C J (2011) Language testing and validation, in O'Sullivan, B (Ed) *Language Testing: Theory & Practice*, Oxford: Palgrave, 13–32.

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

Taylor, L (Ed) (2011) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press.

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

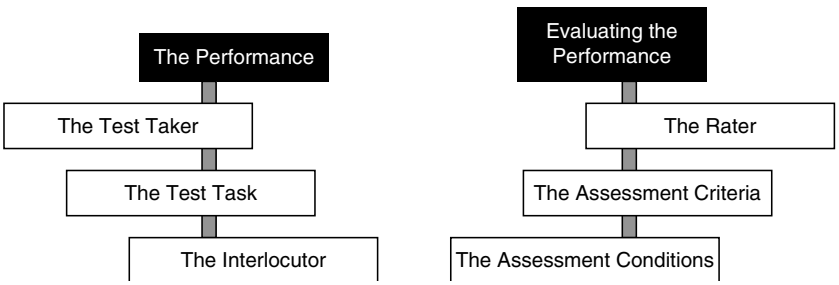
2 Research issues in testing spoken language

Barry O'Sullivan
Cyril J Weir

Section 1: Introduction

This overview of the testing of speaking is intended to offer to the reader our perspective on the central issues involved, and also to present a framework through which the research to date can be evaluated and future research in the area might be formulated. For the purposes of our overview a model of test development is proposed in Figure 1.

Figure 1 Basic organisational framework for this overview



While the elements are displayed as separate entities, it should be noted that particularly within each of the sub sections (performance and evaluating the performance) there will be a degree of interaction between the characteristics. In addition, there may be some interaction across the subsections – for example the rater typically plays the role of interlocutor and rater/examiner in the one-to-one interview.

The oral proficiency of language learners can be tested in a number of ways that involve the candidate participating either in live or simulated interactions. Though other less direct attempts to test oral performance or 'communication' can still be found, such as multiple-choice question (MCQ) pencil-and-paper tests, these indirect tests will not be discussed here as they have very little to tell us about spoken language performance.

The 'live' testing method generally involves what has become known as an oral proficiency interview (OPI) in which the candidate interacts in the target language, either with an interviewer or with another candidate (or both). While in its earlier inception the oral ability of language learners was assessed during a 'live' interview, in which the learner interacted with a lone examiner (resulting in an 'interview' type discourse being produced), this situation has changed over the years. It is now commonly accepted (Lazaraton 1992, 1996a, 1996b, Shohamy 1983) that different test formats or interaction types (such as role-plays, discussions, presentations etc.) are necessary for construct validity. Accordingly, in many current oral tests the candidate engages in a number of different interaction types, which offer a broader view of their overall speaking ability (and hence a more valid assessment of a learner's ability to use the language in terms of both content coverage and theory-based validities). [*Note: the term theory-based validity was the original term Cyril used to describe what he later came to term cognitive validity – in some ways I prefer the original term as it reflects a concept of the construct as an operationalised model of language ability, e.g. in the Khalifa and Weir (2009) reading model. I should also say that the use of the term validities here, and elsewhere in Cyril's 2005 book was never intended to be seen as a rejection of Messick's unitary model; in fact, he always saw the socio-cognitive model as representing a true reflection of the ideas of Messick, whose work he hugely admired.*]

The other commonly used method currently employed is the SOPI, or simulated OPI (also referred to as a tape-mediated test). Here candidate responses to taped prompts are recorded for later evaluation (see Clarke 1979, Lowe and Clifford 1980, and Stansfield and Kenyon 1992 for a fuller description and discussion of the SOPI). The questions employed in this test type tend to be more closed in nature, in that they are designed to elicit responses that can be more readily predicted by the test maker.

OPI has, by and large, continued to dominate oral language proficiency testing due primarily to its apparent face validity (though see van Lier 1989 and McNamara 1996, who question this notion of 'self-evident' validity) and for its relative simplicity in terms of administration. Due to this popularity, the OPI has been receiving a growing amount of interest among language testers and researchers.

We hope it will become clear from our discussion of this work that while the growing interest in research into the testing of speaking has resulted in a corresponding growth in the published work in the area, much of this work has been conducted in the equivalent of a theoretical vacuum. What this report aims to do, therefore, is to provide a practical and theoretical framework upon which future research can be formulated.

Since the testing of spoken language is exclusively performance-based, research might be expected to focus on factors that systematically affect that performance, with an additional focus on other factors that affect test

outcomes and uses. O'Sullivan, Porter and Weir (1999) refined the concept of task difficulty in relation to speaking and three primary areas of concern were highlighted:

- the test taker
- the task
- the examiner/interlocutor.

We will now explore these in detail in Sections 2–4 and then briefly consider assessment criteria (Section 5) and the wider societal context of language testing (Section 6) as additional factors that might influence test outcomes and uses.

Section 2: The test taker

Though the test taker is central to the validation process, the literature relating to characteristics of the test taker is surprisingly small. Those studies that exist have tended to consist of efforts to establish a framework for describing the test taker (a pre-theoretical approach) or efforts to identify particular characteristics of the test taker which affect performance (a pseudo-empirical approach).

The dearth of empirical studies in this area reflects its complexity and the difficulty of dealing with the multiple characteristics involved. However, the studies referred to in Table 1 indicate that, with more complex computer programmes now becoming available – and the corresponding increase in the availability and power of desktop/personal computers – these complexities are becoming less and less problematic. Certainly in the case of tests of speaking, greater attention to design is required if we are to unpack this extremely important but complex area.

Table 1 Summary of research relating to the test taker

Focus	Gloss	See	Issues and questions
Classification of variables	In which writers have attempted to categorise/list those characteristics that are most likely to affect performance. To date these are anecdotal and have not been explored empirically.	Cohen (1994) Bachman and Palmer (1996) Brown (1995) O'Sullivan (2000b, 2002)	➢ Identify characteristics of the test taker which may affect performance – actual and measured ➢ Identify relevant characteristics ➢ Quantify effects
Relationship between variables and performance	In which the researcher has attempted to highlight a variable (or variables) and investigate performance relative to these variables.	Berry (1994, 1997) Kunnan (1995) Purpura (1998)	➢ Explore relationship between quantifiable characteristics identified in the framework and test outcomes (scored performance)

Table 1 (continued)

Focus	Gloss	See	Issues and questions
Interviewee language	In which the primary focus of study was the language of the interviewee or test candidate. These studies, by their nature, have looked at interviewee language in terms of other variables and have been qualitative in nature.	O'Sullivan (2000b) Young and Milanovic (1992)	➢ Task (expected outcomes) ➢ Interlocutor (conversation analysis, discourse analysis) ➢ Examiner/assessment criteria (actual language vs scored performance)

Influenced by the work referred to in the table, but in particular by the work of Brown (1995), O'Sullivan (2000a) suggests that characteristics identified in these studies can be presented as shown in Figure 2.

Figure 2 Characteristics of the test taker

Physical/Physiological	Psychological	Experiential
Short-term ailments *Toothache, cold etc.* Longer term disabilities *Speaking, hearing, vision* Age Sex	Personality Memory Cognitive Style Affective Schemata Concentration Motivation Emotional State	Education Examination Preparedness Examination Experience Communication Experience TL-Country Residence
Language Ability		
Strategic Competence		
Linguistic/Ethnic Background		

In this figure, physical/physiological characteristics can be seen in terms of:

- short-term ailments, such as a toothache or earache, a cold or flu etc. – by their nature these illnesses are unpredictable and are not normally relevant to the construct
- longer-term illnesses or disabilities, such as problems with hearing, vision or speaking – these are predictable and should be taken into account in development and delivery.

The characteristics listed under 'psychological' are ordered to suggest that there will be some that are unlikely to change to any great extent with time,

while others will be more or less likely to change within particular individuals (this list represents an admittedly anecdotally derived continuum).

Experiential characteristics are seen as being comprised of all those influences that have essentially come from outside of the test taker, and refer to their:

- experience of the examination in question – in terms of having prepared through a course of study for example, or having taken the examination previously
- experience in communicating with others, particularly in the target language (TL), but may also refer to L1 communication – this would be of particular concern where, for example, younger learners are expected to interact in the TL with a partner who is unknown to them, something they may rarely have done in their own language
- TL country experience – it is more likely that a learner will experience reduced anxiety having lived for some period of time in the TL country or culture (this distinction refers to the situation where there is a TL subculture within the culture of the learner).

The final three characteristics, language ability, strategic competence and linguistic/ethnic background, are not uniquely categorisable and are seen here as relating to aspects of all three of the superordinate categorisations.

In the above representation, while these characteristics appear to be independent of one another, this clearly is not always the situation. Take, for example, the notion of 'strategy use', seen as an aspect of strategic competence by Bachman (1990). It is likely that the successful (or unsuccessful) use of a particular strategy reflects either a learner's underlying strategic competence (a psychological characteristic), or some experience they may have had of observing successful (or unsuccessful) use of that strategy (an experiential characteristic), or a combination of both.

Section 3: The task

In this report we refer to the conceptualisation of task-based approach offered by Skehan (1996:38) in which he saw the task as an activity, related to a real-world activity and where meaning and task fulfilment are important. Norris, Brown, Hudson and Yoshioka (1998:44, quoting Crookes 1986:27) argue that we need to consider two different parameters for task classification. The first of these, 'assumes tasks can be described and differentiated in terms of the intrinsic objective properties they possess (such as "goals, input stimuli, procedures, responses and stimulus-response relationships")'. The second parameter addresses the 'ability requirements (characteristics of the operator) necessary for tasks varying in complexity on a number of factors'. O'Sullivan (2000a) further refined the first of these parameters by

focusing on individual test taker characteristics and on characteristics of the interlocutor/examiner that are likely to affect performance. To the extent that the variables identified, either singly or in combination, might make task performance easier or more difficult, they obviously impact on the difficulty of spoken language tasks in operation. We will therefore need to examine how interlocutor variables impact on task difficulty but first we will examine the literature specifically relating to the task *per se*. We will investigate the task in terms of:

- differences in **inter-task** levels of difficulty based on empirical research
- **intra-task** characteristics that may vary within the task itself and which research suggests would accordingly make the task more or less difficult.

We take as our starting point the need to determine what empirical evidence is available to support the view that performance will differ from task type to task type and then see how by modifying variables within a single task the difficulty of the task might be altered, i.e. both **inter-** and **intra-task** dimensions. A number of studies in these two dimensions are summarised in Table 2.

It is important to note that, although we discuss the test task itself, the three central factors (test taker, task and examiner/interlocutor) are unlikely to operate in isolation and some degree of interaction is to be expected.

Table 2 Qualitative and quantitative approaches to the dimensions of inter- and intra-task difficulty

Focus	Gloss	See	Issues and questions
Inter-task comparison (Quantitative)	Involving quantitative studies in which comparisons are made between performances on different tasks.	Chalhoub-Deville (1995) Fulcher (1994) Fulcher (1996) Henning (1983) Lumley and O'Sullivan (1999) Norris et al (1998) O'Loughlin (1995) Robinson (1995) Shohamy (1983) Shohamy, Reves and Bejarano (1986) Skehan (1996, 1998) Stansfield and Kenyon (1992) Upshur and Turner (1999) Wigglesworth and O'Loughlin (1993)	How do particular tasks affect a candidate's performance?

Table 2 (continued)

Focus	Gloss	See	Issues and questions
Inter-task comparison (Qualitative)	As above but where qualitative methods are employed.	Bygate (1999) Kormos (1999) O'Sullivan, Weir and Saville (2002) Shohamy (1994) Young (1995)	What differences in language result from using different tasks?
Intra-task performance conditions	Where internal aspects of one task are systematically manipulated (e.g. planning time, pre- or post-task operations etc.).	Brown, Anderson, Shilcock and Yule (1984) Foster and Skehan (1996, 1999) Mehnert (1998) Norris et al (1998) Ortega (1999) Robinson (1995) Skehan (1996, 1998) Wigglesworth (1997)	How will changes to task conditions (e.g. in planning time) affect performance levels on specified criteria?

Inter-task comparisons (quantitative and qualitative)

There have been a number of recent studies that have looked at variability in performance on different test methods. Essentially these can be seen as either 'live' tests – in which the test taker is required to interact with another person (either an examiner or another test taker) – or tape-mediated – where the test taker is required to respond to a series of prompts, delivered by either audio or video tape, and where recordings of these responses are later marked by independent raters.

The only clear generalisation one can make is that, in the majority of the studies, tasks appear to differ in the average levels of performance elicited. However, the research is not sufficiently well grounded to offer clear recommendations on the sequencing of tasks within spoken tests. We summarise in Table 3 the nature and findings of the reported results in this area.

Table 3 Task operation: summary of quantitative differences

Study	Focus	Findings
Brown et al (1984)	Various task design features, in an empirical attempt to establish task difficulty	Static tasks (e.g. description) easier than dynamic tasks (e.g. narration), which in turn are easier than abstract tasks (e.g. opinion giving). The number of elements, participants, and relationships in a task makes it more difficult.

Table 3 (continued)

Study	Focus	Findings
Chalhoub-Deville (1995)	Three tasks (OPI, narration, read-aloud)	Different dimensions appear to underlie the different tasks.
Fulcher (1996)	Group oral results, picture description and discussion	'While task does have a significant effect upon scores, this effect is so small that it does not seriously reduce the ability to generalize from one task to another' (Fulcher 1996:36).
Henning (1983)	Foreign Service Institute (FSI), imitation, completion test	No trait exhibited discriminant validity (i.e. a strong method effect eclipsed the postulated relationship among the three traits).
Lumley and O'Sullivan (1999)	Task difficulty prediction	Varying levels of task difficulty, despite attempts at parallel tasks.
O'Loughlin (1995)	SOPI/OPI comparison	Candidates assigned to different level of proficiency on the two formats.
O'Sullivan (2000a)	Interview/paired task comparison	Tendency to achieve higher scores when engaged in the interview (as v pair work). Evidence of a task/format interaction effect.
Shohamy (1983)	American Council on the Teaching of Foreign Languages (ACTFL) interview/ Reporting article content comparison	Reporting task significantly lower overall than interview tasks (for vocabulary, grammar, and fluency – not significant for pronunciation).
Shohamy et al (1986)	OPI, role-play, reporting, group discussion	Correlations ranged from 0.6 to 0.7 (OPI highest with others, discussion lowest) – as these are low authors suggest the four methods are tapping 'different aspects of oral proficiency'.
Skehan and Foster (1997)	Personal, narrative, and decision-making tasks	Personal task generated less complex language than the narrative and decision-making tasks. Narrative generated the lowest level of accuracy (other two very similar). Personal task highest for fluency (other two tasks broadly similar).
Stansfield and Kenyon (1992)	OPI/SOPI comparison	SOPI test-retest reliability equals or exceeds OPI. SOPI/OPI correlations very high. Generalisability study suggests that variance due to test method is minimal (in all cases most variance is due to 'subjects').
Upshur and Turner (1999)	Personal information exchange (PIE) and story retelling	Candidates' higher performance on PIE.
Wigglesworth and O'Loughlin (1993)	SOPI/OPI comparison	Candidate ability measures strongly correlated ($r = 0.92$). 12% of candidates received different overall classifications for the two tests.

A number of qualitative studies (Table 4) suggest significant differences in the features of the language elicited between tasks but there is no clear evidence concerning the effect these had on test scores and hence on relative difficulty of tasks. They do however support the argument that different tasks elicit samples of language that are qualitatively different and that the effect of this on scoring needs to be investigated.

Table 4 Task operation: summary of qualitative differences

Study	Focus	Findings
Bygate (1999)	Language produced by students performing two tasks (argumentation and narrative) Two versions of each task	Different types of task resulted in significantly different usage of different types of subordinate clause (relative, adverbial, nominal). Argumentative tasks tend to result in short turns. Narrative tasks tend to result in longer turns.
Kormos (1999)	Task comparison: non-scripted interview; guided role-play; picture description	Conversational interaction more symmetrical in the guided role-play. Suggests that this type of task is best to assess candidate's conversational competence.
Shohamy (1994)	OPI/SOPI comparison	No significant difference between formats for a range of linguistic features. Only significant difference was for paraphrasing (SOPI produced more instances). Differences observed in a range of discourse features.
Young (1995)	Conversational styles in OPIs (three tasks – photo-based discussion; relating written passage to same photos; expressing opinions)	Systematic variation in style in language generated by different tasks. One-to-one interview method too scripted so 'may obscure differences among learners' discourse that are relevant to an assessment of their conversational competence' (Young 1995:36).

Intra-task comparisons (quantitative and qualitative)

It does seem that inter-task comparisons, though useful for overall sequencing of tasks in a battery, are probably not sufficiently delicate in themselves for making much progress on establishing what contributes to task difficulty. If we are to make serious inroads into what makes different tasks more or less difficult than each other we need to tie down further as many of the criterial moderating variables as we can. Future socio-cognitive based studies of operations and performance conditions (see McNamara 1996, O'Sullivan 2000a, Weir 1993) might be able to control for these in a systematic and rigorous

fashion when investigating inter- and intra-task differences along specified parameters and measure the effects of these on performance.

Impact on particular aspects of performance

The following components of cognitive processing in performance on speaking tasks have been identified in Skehan's research (1996, 1998):

- Tasks and performance conditions which direct attentional resources to form and rule. These tasks and conditions may induce either 'risk avoiding' or 'risk-taking' behaviour, yielding variation in measures of accuracy and complexity of language, respectively.
- Tasks and performance conditions which focus attentional resources on meaning and real-time processing, yielding variation in measures of fluency.

Foster and Skehan's (1999:221) descriptions of the dependent variables of fluency, accuracy, and complexity are summarised in Table 5.

Table 5 Dependent variables from Foster and Skehan (1999)

Variable	Description
Fluency	Based on number of formulations, replacements, false starts, repetitions, hesitations, and pauses over one second (all reported per 5 minutes of performance).
Accuracy	Based on proportion of error-free clauses (syntax; morphology; word order).
Complexity	Based on measures of subordination, for example, number of clauses per T-unit or c-unit.*

** Clauses are either a simple independent finite clause or a dependent finite or non-finite clause. A c-unit is defined as each independent utterance providing referential or pragmatic meaning.*

The results of a series of studies undertaken by Skehan and Foster (e.g. Foster and Skehan 1996, 1999, Skehan 1996, 1998, Skehan and Foster 1997) suggest that planning conditions can have a significant effect on the ways that tasks are performed. This effect was particularly strong for complexity and for some aspects of fluency, though there was no linear relationship between accuracy and the different planning conditions. In addition, it was clear that when planning was carried out:

- subjects seemed to prioritise *either* complexity *or* accuracy, but not both
- the performance areas (fluency, accuracy and complexity) appear to be in competition with one another.

These results are consistent with an information processing model in which learners have limited capacities, which they deploy selectively, reflecting whatever performance priorities they have or that the tasks and task conditions support.

We have attempted to summarise the suggestions from the literature survey in Tables 6 and 7. These are fairly crude but robust generalisations that represent the overall findings of the limited research done on these areas while recognising that individual studies or contexts may have deviated from these general patterns.

Table 6 Performance conditions and their general impact on difficulty

Moderator variables	Condition	Difficulty
Code complexity	Increased range in input and output	Increase
	More sources of input	Increase
	Imperfect delivery of input	Increase
Cognitive complexity	Amount of input to be processed increases	Increase
	Availability of input decreases	Increase
	Unfamiliar information increases	Increase
	More organisation of information required	Increase
	As information becomes more abstract less concrete	Increase
	Information requiring transformation increases	Increase
	Information requiring retrieval increases	Decrease
Communicative demand	Time pressure increases	Increase
	Response level increases	Increase
	Scale increases	Increase
	Complexity of task outcome increases	Increase
	Referential complexity increases	Increase
	Higher stakes	Increase
	Degree of reciprocity required increases	Increase
	Production involved	Increase
	Task contains a clear macrostructure	Decrease
	As opportunity for control increases	Decrease

Table 7 Performance conditions and their specific impacts on difficulty

Performance condition	Resulting change	Aspect
Unguided planning, especially first minute	⇑	Accuracy
No planning	⇓	Accuracy
Planning >10 minutes	⇑	Complexity
Guided planning	⇑	Fluency/complexity
Post-task activities	⇑	Accuracy
Planning *per se*	⇑	Trade-off between accuracy and complexity, greater fluency
Solitary planning	⇑	Complexity, fluency and turn length

Table 7 (continued)

Performance condition	Resulting change	Aspect
Teacher-led planning	⇑	Accuracy
Group-based planning	Non-significant	All aspects
Different outcomes	⇑	Complexity
Well-structured tasks	⇑	Accuracy
Familiar material	⇑	Fluency
Online computation involved	⇑	Complexity
More cognitively demanding tasks (narrative and decision-making)	⇑	Complexity
Surprise element	Non-significant	All aspects
Pre-task pressure to conformity in structural choice	⇑	Accuracy

Section 4: The examiner/interlocutor

In this section, the literature relating to various aspects of the behaviour of the examiner or interlocutor in OPI situations will be reviewed under the headings: interviewer language; interlocutor variability; examiner/interlocutor effect; rating and the rater.

Interviewer language

Within this section the literature will be divided into those studies which considered interviewer variability in relation to the performance of the test taker, and those where the primary focus was on the language of the examiner/interviewer.

Table 8 summarises the findings of Brown and Lumley (1997), who investigated the effect on test taker performance of variability in the linguistic behaviour of the examiner in the Occupational English Test (OET), indicating that examiner behaviour can have a significant impact on performance – positive and negative.

Table 8 Features of examiner behaviour which affect test difficulty (Brown and Lumley 1997)

Features which make interview easier	Features which make interview harder
Factual questions	Sarcasm
Linguistic simplification	Interruption
Allowing test takers to control the interaction	Repetition
	Lack of co-operation

In Table 9, we summarise the other relevant studies that focused on rater behaviour. The interesting (and disturbing) conclusion we can draw from this research is that it would appear to be very difficult to systematically deliver any live speaking test without significant monitoring of examiner behaviour from a range of perspectives. The studies also seriously question the dependence on correlation statistics as an indication of examiner/test quality.

Table 9 Studies on examiner behaviour

Study	Focus	Findings
Brown (1998)	Analysis of interaction	Structure of sequences of 'talk on topic'; questioning techniques; integration of candidate-nominated topic; provision of feedback. Male interviewers demonstrate far less supportive style.
Young and Milanovic (1992)	Interviewer/test taker interaction – focus on discourse variation (interactional contingency, goal orientation, dominance)	Examiner and candidate contribute differently. Examiner exercises control over the discourse. Different behaviour from male and female interviewers. Interviewee responses dependent on question/task type. Link between theme and topic persistence. Task exerts the strongest influence on the discourse.
Lazaraton (1996a)	Interlocutor/interviewer support	Examiners offered support in a number of different ways, including: • priming topics • supplying vocabulary or collaborative completions • question repetition (slowed speech, more pausing, over-articulation) • drawing conclusions for candidates.
Lazaraton (1996b)	Interlocutor frame adherence	Considerable variation in how individual prompts are used. Systematic variability in the use of particular prompts by the individual examiners.
Ross and Berwick (1992)	Interviewer accommodation	Interviewers adjusted their language to 'facilitate the communication of information during the process of the interview' (Ross and Berwick 1992:164).

Table 9 (continued)

Study	Focus	Findings
Ross (1992)	Use of accommodative questions in the OPI	Factors contributing to the occurrence of accommodation include: • interviewee's response to the previous interviewer question (and accommodation required/ offered) • examiner's perceived level of the interviewee.
Reed and Halleck (1997)	Relationship between interviewer behaviour and test taker performance using the ACTFL guidelines	Unacceptable differences in behaviour in pitch of the line of questioning (an issue as initial pitch level typically determined the final outcome). Highlighted the failure of correlation statistics to satisfactorily reflect interview reliability, suggesting instead that pass/fail distinctions be examined.

Interlocutor variability

Before reviewing the literature on the effect of variables associated with the interlocutor on test performance – both on the actual language elicited and on the scores awarded for the performance – it should first be pointed out that the interlocutor can be either an examiner or another test taker. In the studies reviewed in Table 10, the interlocutor was seen as the former where the focus is underlined.

Table 10 Studies focusing on the effect of features of the interlocutor on performance

Study	Focus	Participants	Findings
Locke (1984)	Interviewer gender	Arab speakers (4 M*)	Better performance with male interviewers.
Porter (1991a, 1991b)	Interviewer gender	Arab speakers (2 F, 11 M)	Better performance with male interviewers.
Porter and Shen (1991)	Gender, status, and interaction style	Mixed nationality (14 F, 14 M)	Better performance with male interviewers. Better performance when interviewer not seen as higher status.

Table 10 (continued)

Study	Focus	Participants	Findings
Berry (1997)	Personality	Hong Kong Chinese (22 F, 32 M)	Females better when partnered with introvert. Men better when partnered with extrovert.
Iwashita (1997)	Proficiency level	Mixed group (20)	No significant difference.
Buckingham (1997)	Gender and age of the interviewer	Japanese (16 F, 16 M)	Female best with female interviewer. Male best with male interviewer.
O'Sullivan and Porter (1996)	Gender of interviewer	Japanese (6 F, 6 M)	Better performance with female interviewers.
O'Sullivan (1995)**	Age of interviewer (in one-to-one) and of peer (in paired task)	Japanese and Arab	Age effect for the older Arab test takers.
Lumley and O'Sullivan (2000)	Task bias (F v M)	Hong Kong Chinese university students (900)	Little evidence of bias effect. Only one task significant (M-oriented task biased towards M).

* *M = Male, F = Female*
** *This bridges the interviewer and peer studies.*

Taken together, the results of these studies indicate that there appears to be sufficient evidence to support a 'sex-of-interlocutor' effect on performance. Of the variables studied to date, this appears to be the one which offers most in terms of significant differences in performance among learners from different cultural backgrounds. Of the other variables, there is, as yet, little empirical evidence of their effect on performance. However, the existing evidence suggests that these variables are worthy of continued exploration, particularly in light of the other studies reported on in this review, which indicate that test performance (and the scores awarded for that performance) is affected by a range of variables.

In the area of second language acquisition (SLA) there have been a number of studies which focused on language variation, where the emphasis has been on variation related to performance conditions. Ellis (1989:22) began his study on intra-learner variability in language use with the statement: 'The existence of variability within the second language use of a single learner is acknowledged by all researchers, irrespective of their theoretical standpoint.' This variability has been discussed in terms of its relation to the *topic* on which the learners are asked to speak (Cornu and Delahaye 1987, Dowd

1984, Selinker and Douglas 1985, Smith 1989, Zuengler 1982, 1989, 1993, Zuengler and Bent 1991); on the *ethnicity* of the interlocutor (Beebe 1977, Beebe and Zuengler 1983, Young 1987); on the *relative status* of the interlocutors (Thakerar, Giles and Cheshire 1982 – with native speaker–native speaker (NS–NS) dyads; and Zuengler 1989 – with native speaker–non-native speaker (NS–NNS) dyads); and on the *relative sex* of the interlocutors (Gass and Varonis 1986, Leet-Pellegrini 1980). In addition, there have been a number of studies devoted to the exploration of the concept of dominance in interactions, not always unrelated to some of the above factors (Scarcella 1983 – with NNS performances; and others such as Ferguson 1977 and Zimmerman and West 1975 – with NS performances); for a more complete review of this literature see Tarone (1988).

These studies suggest that the interlanguage of any language learner who is involved in a communicative interaction with an interlocutor, be they a student in an EFL classroom or a test taker participating in a test of speaking ability, will be affected by a number of factors, including the situation or context, the topic, and the person with whom the learner is interacting.

Examiner/interlocutor effect

In this part we return to the examiner/interlocutor effect discussed earlier. However, here the focus of interest is the effect on task performance of variability in a number of features of the interlocutor. This was the subject of O'Sullivan's (2000a) thesis, in which he investigated a number of these variables and demonstrated how the variables interact with one another to affect performance (see Table 11).

Table 11 Studies focusing on the effect of features of the interlocutor on performance

Variable	Findings
Gender	Candidates tend to achieve higher scores when interacting with a female interviewer – though culturally dependent (higher for Japanese, lower for Arab).
Age	No evidence that this is an issue in a test context.
Language level: interlocutor's actual or perceived linguistic ability	Where test takers of mixed ability are paired, we might expect that the weaker student would gain in terms of support during the activity, but would tend to cede the floor to their partner, while the *more able* student may be held back to some degree by their *weaker* partner.
Acquaintanceship	Where a candidate is paired with a person considered to be a friend, they will be expected to perform better than when their partner is a stranger.
Personality	Some evidence (Berry 1997) of an effect in paired/group tests. No evidence of a main effect in terms of individual test taker's perception of relative personality type of interlocutor.

Table 11 (continued)

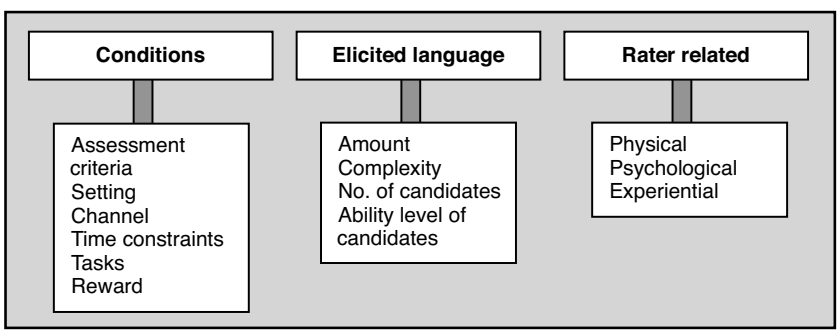
Variable	Findings
Interaction between factors	O'Sullivan (2000a) supports the view that particular factors (such as sex, relative age and acquaintanceship) can, when isolated under experimental conditions, significantly and systematically affect performance. Significant interaction effect indicating sufficient evidence that test takers' reactions to their partner affect their performances on tasks.

There had been a small number of studies in which more than one of the factors have been included (see Berry 1997, O'Sullivan 2000a, O'Sullivan and Porter 1997 in EFL, and Feeney and Kirkpatrick 1996 in psychology), the results of which highlight the complex interaction of all these variables when studied in affecting performance.

Rating and the rater

When we consider the area of the scoring (or rating) of test performance we must include in our theoretical conceptualisation a whole series of related factors. This broad perspective might be expressed using the overview suggested in Figure 3.

Figure 3 Conceptual overview of factors affecting the rating process



These factors are explored in more detail in the following section discussing assessment criteria.

Section 5: Assessment criteria

Alderson (1991) refers to a scale as having three potential purposes:

- user-oriented (to enable users to interpret the results – reporting function)

- assessor-oriented (to guide assessors who evaluate the performance)
- constructor-oriented (to provide guidelines to test constructors – test construction function).

A number of methods for scale construction are suggested: expert/stakeholder evaluation of draft descriptions; experienced markers identifying performances at different levels then agreeing on 'key' features; discussion leading to identification of criteria for assessment defined in terms of performance levels; and fine-tuning from feedback data obtained through trials and usage – a continuous process. Table 12 offers a summary of the research to date.

Table 12 Scale construction studies

Focus	Gloss	See	Issues and questions
Scale construction 1: Theoretical/ validity issues	Where the criteria included in a scale are validated against 'objective' measures; includes issues of assumptions of 'linearity' in scales.	Abdul Raof (2002) Alderson (1991) Fulcher (1994, 1996) North (1995) North and Schneider (1998) Pollitt and Murray (1996) Upshur and Turner (1995)	➢ Do scales reflect an acceptable theory of language proficiency? ➢ Is 'actual language use' accurately measured by the existing scales? ➢ Do scales operate similarly at different proficiency levels? ➢ Should we report a score or a profile?
Scale construction 2: Practicality/usage issues	Includes issues such as number of criteria; planned/ unplanned emphasis on criteria; degree of ease or difficulty in application of criteria.	Chambers and Richards (1996) Fulcher (1996) Lumley and McNamara (1995)	➢ Comparability of the holistic and analytic versions of scales ➢ Individual-task assessment vs. global performance assessment ➢ Does rater training affect scale interpretation?
Scale comparison	Where comparisons have been made of measures awarded for performance on two separate scales.	Hasselgren (1997) O'Sullivan (1995)	➢ Can rater usage of the different scale types (analytic/ holistic) be compared? ➢ What are the scales telling us? (Different/same information?)

Looking across these studies, it is clear that there is no easy answer to the question of which approach is best. Clearly, the answer to the question is 'it depends' on the test, the purpose, and the context.

Setting

Typical rating settings include:

- 'live' rating of performances at the test venue – familiar or unfamiliar to the rater
- raters visiting an examination board to award scores to recordings of test events
- raters receiving recordings of test events to score at a venue of their own choice.

It seems clear that variation in the above settings may lead to systematic variability in the scores awarded in the rating process. To date however, there has been no empirical research into the effect on rater performance of factors associated with the setting for the rating process.

Channel

Some evidence exists that there are differences in the ratings awarded depending on channel (O'Loughlin 1995, Thompson 1995), while McNamara and Lumley (1997) found that that raters were influenced by factors other than language performance (e.g. sound quality) when awarding scores to test takers from audio-recording of test events.

Time constraints

As with the references to the physical conditions of the rating procedure, there is a real possibility that temporal factors may also influence rating performance. This effect is most obviously seen in the 'live' versus recorded dichotomy. One cause of difficulty in the 'live' version of a test is the ephemeral nature of language production – it happens and is gone. The rater must make 'real-time' judgements of the test taker's ability with no second chance. On the other hand, the rater working with a taped/recorded performance has the opportunity to pause, stop or replay the tape at any time. This security may (even where not used) affect the way they behave during the rating procedure.

Another source of variability in the rating process is the pressure on raters to award scores during an actual performance. While it may benefit the reliability of the rating process to have multiple scores for each performance (e.g. a single score per task where test takers perform a series of tasks), the added

time pressure on the rater (particularly as there is often a seamless move between tasks) may in fact negate any benefit accrued.

Reward

A final area of concern here is that of the motivation of the rater. While there has been some work done on the effect of motivation on the test performance of test takers, there has not been any published work to date which focuses on the rater.

Elicited language

The second major influence on the rating procedure is related to the actual language elicited during the test event. While there is some connection with the previous section here, in that the clarity of the language can depend on one or more of the conditions referred to previously (most notably channel and setting), the main thrust of this brief section is to highlight other aspects of the elicited language that are more related to the individual test events than to the rating conditions. The impact (or potential impact) of a range of aspects of language is summarised in Table 13.

Table 13 Impact of aspects of candidate language on rating

Aspect	Evidence
Amount	It is conceivable that a test taker who regularly expands on responses to examiner prompts, or who, in a group or pair task consistently opens turns, will benefit when it comes to the rating process. It would therefore be of real interest to explore how raters might be influenced by the sheer size of a test taker's output, whether this is related to an overall performance or to specific aspects of a performance such as those suggested above.
Complexity	When we refer to complexity here, we have in mind the complexity of the content or the ideas of the response rather than any complexity of the language used to create that response. As this is not typically included in rating scale descriptors, it may be an issue worth investigating.
Number of candidates	The number of candidates being tested at any one time may also affect the scores awarded, either through the rater becoming overloaded in attempting to monitor the production of all candidates in a group, for example, or through the rater making conscious or subconscious allowances for non/over participation.
Ability level of candidates	While we expect that the ability level of the test takers will affect the scores awarded (if they do not we have a *real* problem), it is not clear whether raters behave differently with candidates of different levels. Among the possible questions to be answered here are: • Which level is easiest to score consistently? • Is it easier to score consistently where the test takers are at approximately the same level?

Rater related

What makes a 'good' rater? We need to be aware of the fact that the rating process is a performance in its own right, and as such is likely to be affected by a number of variables. These variables are either related to the rater, to the task or to the candidate (or more precisely, to the rater's affective reaction to the task and/or the candidate) or to the test conditions (which is one reason why we specify the conditions so carefully). O'Sullivan (2000a:19) suggests that we need to treat the examiner/rater in much the same way as we think about the test taker (see Table 14).

Table 14 Rater characteristics to be considered

Aspect	Evidence
Physical/Physiological	While there is no evidence that raters of a particular gender make 'better' raters, it is conceivable that under specific circumstances there may be an interaction between the 'gender' variable and some other element of a test (e.g. a particular task, or in an extreme case a particular body of test takers). Another potential effect of these physical characteristics of the rater is the possibility that certain candidates may react in either a positive or a negative way to them. This is particularly relevant in tests of speaking where the rater plays the role of examiner and interlocutor.
Psychological	There is no reason why an affective reaction on the part of the test taker to perceptions of the physical – and of course psychological – characteristics of the rater/examiner cannot be mutual, i.e. that an examiner may be unduly influenced by characteristics of the test taker. Wade and Kiniki (1997:35) found evidence that interviewers' subjective impression completely mediates the relationship between the interviewer and the interviewee.
Experiential	There have been a small number of studies which have looked at rater background and experience of the rater/examiner, see Table 15.

Examiner/rater experience and training

Though examination experience can refer to a whole range of experiential characteristics, such as knowledge of the test, knowledge and experience of the role adapted (examiner, interlocutor, facilitator) or experience of examining at the level of the test, most research has tended to focus on the effects of rater training on rater performance.

The importance of rater training has been stressed in the literature (Alderson, Clapham and Wall 1995:112, Bachman and Palmer 1996:222, Brown 1995:2–3, Weir 1988:89), and has been the subject of a number of

studies which have explored rater training from a number of perspectives. These are summarised in Table 15.

Table 15 Examiner experience studies

Study	Focus	Findings
Magnan (1987)	Inter-rater reliability of trainee ACTFL examiners	High degree of agreement (however, one in five were deemed unacceptable plus high numbers of 'unrateable' performances were still rated by trainees in contravention to their training).
Wigglesworth (1993)	Performance in a rating procedure	It is possible to reduce rater bias during rating.
Brown (1995)	Rater background (occupation and language)	Considerable differences in individual degrees of reliability, bias and candidate ranking. No evidence of significant differences between the different groups.
Chalhoub-Deville (1995)	Differences in rater behaviour	Different rater groups appear to focus on different dimensions for different tasks.
Halleck (1996)	Inter-rater reliability of novice ACTFL examiners	Little difficulty with superior- and intermediate mid-level performances. Mixed results (one perfect, one mixed, one problematic).
Kenyon (1997)	Self-instructed rater training in the context of the SOPI	Self-instruction at least 'appears possible'.
Weigle (1998)	Long-term impact of training	Need for rater training beyond the initial standardisation process.
Lumley (1998)	Language-trained vs subject-specialist raters	Broad agreement between the groups at a global level.
Upshur and Turner (1999)	Investment in scale	Raters who had some investment in the scale applied it more harshly. Raters tended to agree on the PIE task scores but varied widely in terms of harshness for the story-retelling task.
Lumley and O'Sullivan (2000)	Experience and rating performance	Tendency for newly trained raters to be harsh, though consistent. Experienced raters tended to have become more lenient; some were also prone to inconsistency.
O'Sullivan and Rignall (2002)	Effect on rater performance of systematic feedback	Not as clear cut as Wigglesworth – did not result in more consistent scoring.

Section 6: The societal context

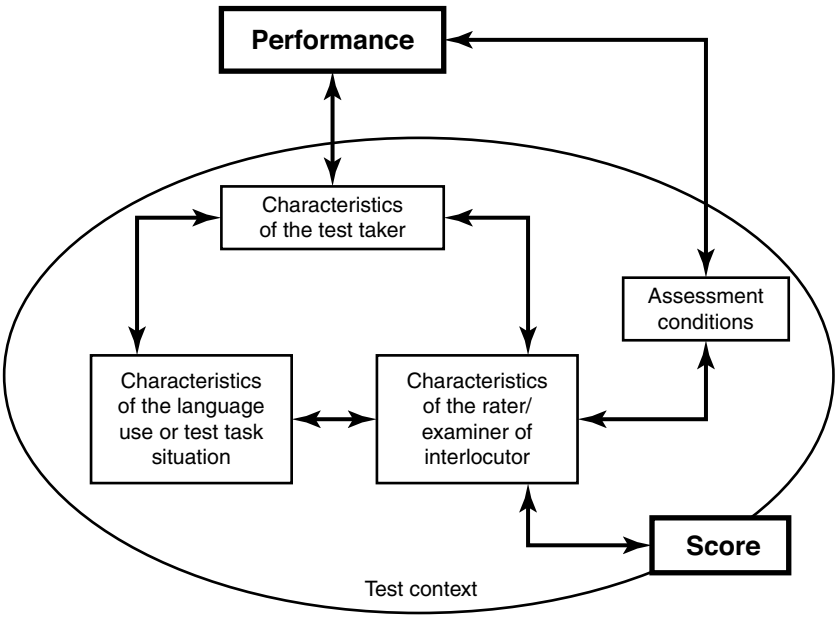
Testing spoken language obviously does not take place in a social vacuum and though not directly relevant to the particular focus of this review on factors that affect actual test performance on spoken language tasks, the wider societal context should not be ignored. The context of the situation in which tests are developed, administered and their results used is an area of research that is receiving increased attention in the literature. Professional accountability, the development of explicit standards for test makers, upon which their tests are to be based, is increasingly recognised as a s*ine qua non* of test development. Ethical accountability, which entails considering how tests and the results of tests are used and the responsibilities of the test maker to established stakeholders, is also receiving due consideration as researchers attempt to gather empirical evidence of test impact/washback and usage (Hamp-Lyons 1997, Rea-Dickins 1997). Legal accountability, which acknowledges the formal/legal responsibilities of test makers to test takers and end users, is also becoming recognised (Fulcher and Bamford 1997). Finally, as part of a new approach to test fairness conditions (Kunnan 2000), the issue of test access and in particular the area of test accommodations for disadvantaged candidates can no longer be ignored by examination boards.

The way these wider societal issues impact on the actual performance of candidates in a test has yet to be researched. Like so many areas in this overview we can only hope the reader is motivated to explore them with us in the future.

Section 7: Modelling the speaking test event

Taking the work of O'Sullivan (2000a) one step further towards a model of the overall test event, the overview of the issues highlighted in this chapter suggests that the speaking test event cannot be modelled as simply as has been suggested in earlier literature. Instead, any model must take account of the notion of 'affect'. We have seen from O'Sullivan's model of performance, which has been supported empirically, that any model will have to include an acknowledgement of the interactions that take place between the variables included. [*Note: the concept of affect reflects my research into affective factors in speaking test performance – in fact, this is where the whole idea of socio came from as this research seemed to be telling me/us that the cognitive load on task performance was affected by social variables such as age, gender and perception of attribute associated with the interlocutor – the wording affective factors came from Don Porter from his earlier work in the early 1990s.*]

Figure 4 A working model of the speaking test event

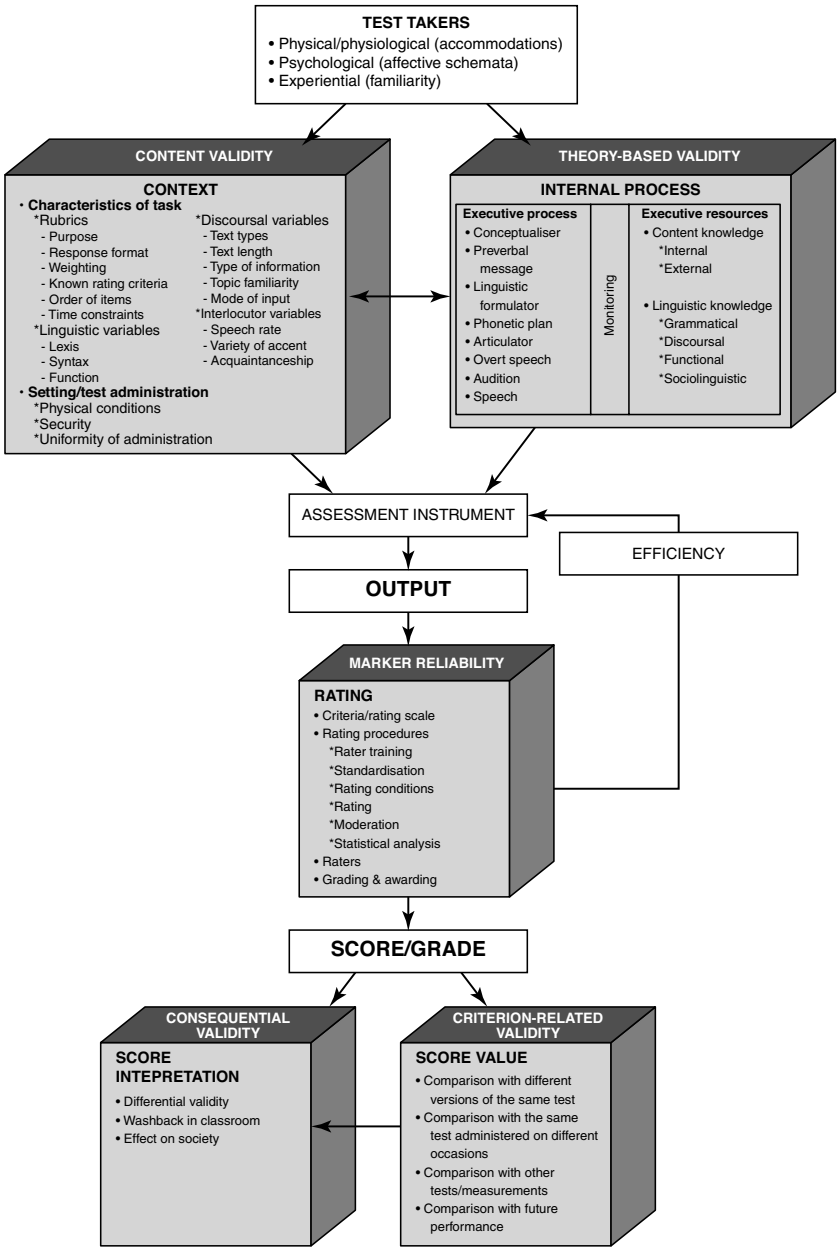


In our model (Figure 4), the performance aspect has been slightly modified from O'Sullivan's original conceptualisation, to highlight the direct relationship between the test taker and the performance (in the original, this relationship was not made as explicit). The interactions between the three sets of variables (i.e. test taker, task, examiner/interlocutor) seen to be associated with performance are now viewed within the context of the broader test event, thus acknowledging that these relationships are test related, and may not necessarily be generalisable outside of the event.

Similarly, the assessment conditions are also part of the event and will include things like the selection criteria, training procedures, accreditation procedures, support (such as feedback), the physical conditions under which rating is performed and the assessment criteria. Finally, the score is awarded by the rater, but having first been affected by the interaction of the many variables suggested in the model. The score is also an artefact of the test event, but is shown as also having an impact beyond that context into a wider societal context.

Finally, in Figure 5 we present a preliminary framework which summarises the thinking which drives this chapter. It builds on the earlier work of Weir (1993) as well as on his current thinking on the importance of cognition in test performance, and on O'Sullivan's more recent work on the effect on test performance of the social context of language use. This framework is meant to

Figure 5 A preliminary socio-cognitive framework for testing speaking



offer test developers a sound basis for both development and validation in the future. [*Note: while there are some differences in the framework to the version presented in Cyril's 2005 book, you can see that these are essentially limited to the naming conventions – some aspects were slow to change – e.g. 'theory-based' eventually became 'cognitive validity' and then 'test-taker model' – while others did not make it to the next version we used internally – e.g. 'marker reliability' became 'scoring validity', and then 'scoring system'.*]

References

Abdul Raof, A H (2002) *The validity of criteria for the assessment of spoken English*, unpublished PhD dissertation, The University of Reading.

Alderson, J C (1991) Bands and scores, in Alderson, J C and North, B (Eds) *Language Testing in the 1990s*, London: Macmillan, 71–86.

Alderson, J C, Clapham, C and Wall, D (1995) *Language Test Construction and Evaluation*, Cambridge: Cambridge University Press.

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

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

Beebe, L (1977) The influence of the listener on code switching, *Language Learning* 27, 331–339.

Beebe, L and Zuengler, J (1983) Accommodation theory: An explanation for style shifting in second language dialects, in Wolfson, N and Judd, E (Eds) *Sociolinguistics and Language Acquisition*, Rowley: Newbury House, 195–213.

Berry, V (1994) *Personality characteristics and the assessment of spoken language in an academic context*, paper presented at the 16th Annual Language Testing Research Colloquium, Washington, DC, USA, 5–7 March 1994.

Berry, V (1997) *Gender and personality as factors of interlocutor variability in oral performance tests*, paper presented at the 19th Annual Language Testing Research Colloquium, Orlando, Florida, 6–9 March 1997.

Brown, A (1995) The effect of rater variables in the development of an occupation-specific language performance test, *Language Testing* 12 (1), 1–15.

Brown, A (1998) *Interviewer style and candidate performance in the IELTS oral interview*, paper presented at the 20th Language Testing Research Colloquium, Monterey, California, 9–12 March 1998.

Brown, A and Lumley, T (1997) Interviewer variability in specific-purpose language performance tests, in Huhta, A, Kohonen, V, Kurki-Suonio, L and Luoma, S (Eds) *Current Developments and Alternatives in Language Assessment: Proceedings of LTRC 96*, Jyväskylä: University of Jyväskylä Press, 137–150.

Brown, G, Anderson, A, Shilcock, R and Yule, G (1984) *Teaching Talk: Strategies for Production and Assessment*, Cambridge: Cambridge University Press.

Brown, J D (1996) *Testing in Language Programs*, Upper Saddle River: Prentice Hall Regents.

Buckingham, A (1997) *Oral Language Testing: Do the Age, Status and Gender of the Interlocutor Make a Difference*?, unpublished MA dissertation, University of Reading.

Bygate, M (1999) Quality of language and purpose of task: patterns of learners'

language on two oral communication tasks, *Language Teaching Research* 3 (3), 185–214.
Chalhoub-Deville, M (1995) A contextualized approach to describing oral language proficiency, *Language Learning* 45 (2), 251–281.
Chambers, F and Richards, B (1996) Reliability and validity in the GCSE oral examination, *Language Learning Journal* 14, 28–34.
Clarke, J L D (1979) Direct vs. semi-direct tests of speaking ability, in Brière, E J and Hinofotis, F B (Eds) *Concepts in Language Testing: Some Recent Studies*, Washington, DC: TESOL, 35–49.
Cohen, A (1994) *Assessing Language Ability in the Classroom* (Second edition), Boston: Heinle and Heinle.
Cornu, A M and Delahaye, M (1987) Variability in interlanguage reconsidered: LSP vs. non-LSP IL talk, *English for Specific Purposes* 6, 145–152.
Crookes, G V (1986) *Task Classification: A Cross-disciplinary Review*, Technical Report No. 4, Hawai'i: Center for Second Language Research, Social Science Research Institute, University of Hawai'i.
Dowd, J L (1984) *Phonological Variation in L2 Speech: The Effects of Emotional Questions and Field Dependence/Field Independence on Second Language Performance*, unpublished doctoral dissertation, Teachers College, Columbia University.
Ellis, R (1989) Sources of intra-learner variability in language, in Gass, S, Madden, C, Preston, D and Selinker, L (Eds) *Variation in Second Language Acquisition Volume II: Psycholinguistics Issues*, Philadelphia: Multilingual Matters, 22–45.
Feeney, B C and Kirkpatrick, L A (1996) Effects of adult attachment and presence of romantic partners on physiological responses to stress, *Journal of Personality and Social Psychology* 70 (2), 255–270.
Ferguson, N (1977) Simultaneous speech, interruptions and dominance, *Journal of Social and Clinical Psychology* 16, 295–302.
Foster, P and Skehan, P (1996) The influence of planning and task type on second language performance, *Studies in Second Language Acquisition* 18, 299–323.
Foster, P and Skehan, P (1999) The influence of source of planning and focus of planning on task-based performance, *Language Teaching Research* 3 (3), 215–247.
Fulcher, G (1994) Some priority areas for oral language testing, *Language Testing Update* 15, 39–47.
Fulcher, G (1996) Does thick description lead to smart tests? A data-based approach to rating scale construction, *Language Testing* 13 (2), 208–238.
Fulcher, G and Bamford, R (1997) I didn't get the grade I need. Where's my solicitor?, *System* 24 (4), 437–448.
Gass, S M and Varonis, E M (1986) Sex differences in nonnative speaker–nonnative speaker interactions, in Day, R R (Ed) *Talking to Learn*, Rowley: Newbury House, 327–351.
Halleck, G (1996) Interrater reliability of the OPI: Using academic trainee raters, *Foreign Language Annals* 29 (2), 223–238.
Hamp-Lyons, L (1997) Washback, impact and validity: ethical concerns, *Language Testing* 14 (3), 295–303.
Hasselgren, A (1997) Oral test subskill scores: what they tell us about raters and pupils, in Huhta, A, Kohonen, V, Kurki-Suonio, L and Luoma, S (Eds) *Current Developments and Alternatives in Language Assessment: Proceedings of LTRC 96*, Jyväskylä: University of Jyväskylä Press, 241–256.

Henning, G (1983) Oral proficiency testing: comparative validities of interview, imitation, and completion methods, *Language Learning* 33 (3), 315–332.

Iwashita, N (1997) *The validity of the paired interview format in oral performance testing*, paper presented at the 19th Annual Language Testing Research Colloquium, Orlando, Florida, 6–9 March 1997.

Kenyon, D M (1997) Further research on the efficacy of rater self-training, in Huhta, A, Kohonen, V, Kurki-Suonio, L and Luoma, S (Eds) *Current Developments and Alternatives in Language Assessment: Proceedings of LTRC 96*, Jyväskylä: University of Jyväskylä Press, 257–273.

Kormos, J (1999) Simulating conversations in oral proficiency assessment: a conversation analysis of role plays and non-scripted interviews in language exams, *Language Testing* 16 (2), 163–188.

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

Kunnan, A J (2000) Fairness and justice for all, in Kunnan, A J (Ed) *Validation in Language Assessment: Selected Papers from the 17th Language Testing Research Colloquium*, Mahwah: Lawrence Erlbaum Associates, 1–10.

Lazaraton, A (1992) The structural organisation of a language interview: a conversational analytic perspective, *System* 20 (3), 373–386.

Lazaraton, A (1996a) Interlocutor support in oral proficiency interviews: the case of CASE, *Language Testing* 13 (2), 151–172.

Lazaraton, A (1996b) A qualitative approach to monitoring examiner conduct in CASE, in Milanovic, M and Saville, N (Eds) *Performance Testing, Cognition and Assessment: Selected Papers from the 15th Language Testing Research Colloquium*, Studies in Language Testing volume 3, Cambridge: UCLES/Cambridge University Press, 18–33.

Leet-Pellegrini, H M (1980) Conversational dominance as a function of gender and expertise, in Giles, H, Robinson, W P and Smith, P (Eds) *Language: Social Psychological Perspectives*, New York: Pergamon Press, 97–104.

Locke, C (1984) *The influence of the interviewer on student performance in tests of foreign language oral/aural skills*, unpublished MA project, University of Reading.

Lowe P, Jnr, and Clifford, R T (1980) Developing an indirect measure of overall oral proficiency, in Frith, J R (Ed) *Measuring Spoken Language Proficiency*, Washington, DC: Georgetown University Press, 31–39.

Lumley, T (1998) Perceptions of language-trained raters and occupational experts in a test of occupational English language proficiency, *English for Specific Purposes* 17, 347–367.

Lumley, T and McNamara, T F (1995) Rater characteristics and rater bias: implications for training, *Language Testing* 12 (1), 54–71.

Lumley, T and O'Sullivan, B (1999) *Report of the Final Trials of the Graduating Students Language Proficiency Assessment (GSLPA)*, Hong Kong: The Hong Kong Polytechnic University.

Lumley, T and O'Sullivan, B (2000) *The effect of speaker and topic variables on task performance in a tape-mediated assessment of speaking*, paper presented at the 2nd Annual Asian Language Assessment Research Forum, The Hong Kong Polytechnic University, January 2000.

Magnan, S S (1987) Rater reliability of the ACTFL Oral Proficiency Interview, *Canadian Modern Language Review* 43, 525–537.

McNamara, T F (1996) *Measuring Second Language Performance*, London: Longman.
McNamara, T F and Lumley, T (1997) The effect of interlocutor and assessment mode variables in overseas assessments of speaking in occupational settings, *Language Testing* 14 (2), 140–156.
Mehnert, U (1998) The effects of different lengths of time for planning on second language performance, *Studies in Second Language Acquisition* 20, 83–108.
Norris, J, Brown, J D, Hudson, T and Yoshioka, J (1998) *Designing Second Language Performance Assessments*, Technical Report No. 18, Hawai'i: University of Hawai'i Press.
North, B (1995) *The development of a common framework scale of descriptors of language proficiency based on a theory of measurement*, unpublished PhD dissertation, Thames Valley University.
North, B and Schneider, G (1998) Scaling descriptors for language proficiency scales, *Language Testing* 15 (2), 217–263.
O'Loughlin, K (1995) Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test, *Language Testing* 12 (2), 217–237.
O'Sullivan, B (1995) *Oral Language Testing: Does the Age of the Interlocutor Make a Difference?*, unpublished MA dissertation, University of Reading.
O'Sullivan, B (2000a) *Towards a Model of Performance in Oral Language Testing*, unpublished PhD dissertation, University of Reading.
O'Sullivan, B (2000b) Exploring gender and oral proficiency interview performance, *System* 28 (3), 373–386.
O'Sullivan, B (2002) Learner acquaintanceship and oral proficiency test pair-task performance, *Language Testing* 19 (3), 277–295.
O'Sullivan, B and Porter, D (1996) *Speech Style, Gender and Oral Proficiency Interview Performance*, paper presented at the RELC Conference, Singapore, April 1996.
O'Sullivan, B and Porter, D (1997) *The Effect of Learner Acquaintanceship on Pair-Task Performance*, paper presented at the SEAMEO RELC Conference, Singapore, April 1997.
O'Sullivan, B and Rignal, M (2002) *A Longitudinal Study of the Effects of Feedback on Raters of the IELTS Writing Module*, IELTS Australia/British Council/UCLES funded research paper.
O'Sullivan, B, Porter, D and Weir, C J (1999) *Research Issues in Testing Spoken Language*, internal research report commissioned by UCLES.
O'Sullivan, B, Weir, C J and Saville, N (2002) Using observation checklists to validate speaking-test tasks, *Language Testing* 19 (1), 33–56.
Ortega, L (1999) Planning and focus on form in L2 oral performance, *Studies in Second Language Acquisition* 20, 109–148.
Pollitt, A and Murray, N L (1996) What raters *really* pay attention to, in Milanovic, M and Saville, N (Eds) *Performance Testing, Cognition and Assessment*: *Selected Papers from the 15th Language Testing Research Colloquium*, Studies in Language Testing volume 3, Cambridge: UCLES/ Cambridge University Press, 74–91.
Porter, D (1991a) Affective factors in language testing, in Alderson, J C and North, B (Eds) *Language Testing in the 1990s*, London: Macmillan, 32–40.
Porter, D (1991b) Affective factors in the assessment of oral interaction: gender and status, in Arnivan, S (Ed) *Current Developments in Language Testing*, Singapore: SEAMEO Regional Language Centre, Anthology Series 25, 92–102.

Porter, D and Shen, S H (1991) Sex, status and style in the interview, *The Dolphin* 21, 117–128.
Purpura, J (1998) Investigating the effects of strategy use and second language test performance with high- and low-ability test takers: a structural equation modelling approach, *Language Testing* 15 (3), 333–379.
Rea-Dickins, P (1997) So, why do we need relationships with stakeholders? A view from the UK, *Language Testing* 14 (3), 304–314.
Reed, D and Halleck, G B (1997) Probing the ceiling in oral interviews: what's up there?, in Huhta, A, Kohonen, V, Kurki-Suonio, L and Luoma, S (Eds) *Current Developments and Alternatives in Language Assessment: Proceedings of LTRC 96*, Jyväskylä: University of Jyväskylä Press, 225–238.
Robinson, P (1995) Task complexity and second language narrative discourse, *Language Learning* 45 (1), 99–140.
Ross, S (1992) Accommodative questions in oral proficiency interviews, *Language Testing* 9, 173–186.
Ross, S and Berwick, R (1992) The discourse of accommodation in oral proficiency interviews, *Studies in Second Language Acquisition* 14, 159–176.
Scarcella, R (1983) Discourse accent in second language performance, in Gass, S M and Selinker, L (Eds) *Language Transfer in Language Learning*, Rowley: Newbury House, 306–326.
Selinker, L and Douglas, D (1985) Wrestling with context in interlanguage theory, *Applied Linguistics* 6, 190–204.
Shohamy, E (1983) The stability of oral language proficiency assessment on the oral interview testing procedure, *Language Learning* 33, 527–540.
Shohamy, E (1994) The validity of direct versus semi-direct oral tests, *Language Testing* 11, 99–123.
Shohamy, E, Reves, T and Bejarano, Y (1986) Introducing a new comprehensive test of oral proficiency, *ELT Journal* 40 (3), 212–220.
Skehan, P (1996) A framework for the implementation of task-based instruction, *Applied Linguistics* 17, 38–62.
Skehan, P (1998) *A Cognitive Approach to Language Learning*, Oxford: Oxford University Press.
Skehan, P and Foster, P (1997) The influence of planning and post-task activities on accuracy and complexity in task-based learning, *Language Teaching Research* 1 (3), 185–211.
Smith, J (1989) Topic and variation in ITA oral proficiency, *English for Specific Purposes* 8, 155–168.
Stansfield, C W and Kenyon, D M (1992) Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview, *System* 20, 347–364.
Tarone, E (1988) *Variation in Interlanguage*, London: Edward Arnold.
Thakerar, J, Giles, H and Cheshire, J (1982) Psychological and linguistic parameters of speech accommodation theory, in Fraser, C and Scherer, K (Eds) *Advances in the Social Psychology of Language*, Cambridge: Cambridge University Press, 205–255.
Thompson, I (1995) A study of interrater reliability of the ACTFL Oral Proficiency Interview in five European languages: Data from ESL, French, German, Russian and Spanish, *Foreign Language Annals* 28 (3), 407–422.
Upshur, J A and Turner, C (1995) Constructing rating scales for second language tests, *ELT Journal* 49 (1), 3–12.

Upshur, J A and Turner, C (1999) Systematic effects in the rating of second-language speaking ability: test method and learner discourse, *Language Testing* 16 (1), 82–111.
van Lier, L (1989) Reeling, writhing, drawling, stretching, and fainting in coils: oral proficiency interviews as conversation, *TESOL Quarterly* 23, 489–508.
Wade, K J and Kiniki, A J (1997) Subjective applicant qualifications and interpersonal attraction as mediators within a process model of interview selection decisions, *Journal of Vocational Behaviour* 50, 23–40.
Weigle, S C (1998) Using FACETS to model rater training effects, *Language Testing* 15 (2), 263–287.
Weir, C J (1988) *Communicative Language Testing*, Exeter: Exeter University Press.
Weir, C J (1993) *Understanding and Developing Language Tests*, Hemel Hempstead: Prentice Hall.
Wigglesworth, G (1993) Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction, *Language Testing* 10 (3), 305–336.
Wigglesworth, G (1997) An investigation of planning time and proficiency level on oral test discourse, *Language Testing* 14 (1), 85–106.
Wigglesworth, G and O'Loughlin, K (1993) An investigation into the comparability of direct and semi-direct versions of an oral interaction test in English, *Melbourne Papers in Language Testing* 2 (1), 56–67.
Young, R (1987) *Variation and the interlanguage hypothesis*, paper presented at the TESOL Conference, Miami, Florida, 21–25 April 1987.
Young, R (1995) Conversational styles in language proficiency interviews, *Language Learning* 45 (1), 3–42.
Young, R and Milanovic, M (1992) Discourse variation in oral proficiency interviews, *Studies in Second Language Acquisition* 14, 403–424.
Zimmerman, D H and West, C (1975) Sex roles, interruptions and silences in conversation, in Thorne, B and Henley, N (Eds) *Language and Sex: Difference and Dominance*, Rowley: Newbury House, 105–129.
Zuengler, J (1982) Applying accommodation theory to variable performance data in L2, *Studies in Second Language Acquisition* 4, 181–192.
Zuengler, J (1989) Assessing an interaction-based paradigm: How accommodative should we be?, in Eisenstein, M R (Ed) *The Dynamic Interlanguage*, New York: Springer, 49–67.
Zuengler, J (1993) Explaining NNS interactional behavior: The effect of conversational topic, in Kasper, G and Blum-Kulka, S (Eds) *Interlanguage Pragmatics*, Oxford: Oxford University Press, 184–195.
Zuengler, J and Bent, B (1991) Relative knowledge of content domain: An influence on native non–native conversations, *Applied Linguistics* 12, 397–415.

3 Cyril Weir and cognitive validity

John Field
Centre for Research in English Language Learning and Assessment, Bedfordshire

If we can identify the skills and strategies that appear to make an important contribution to the reading process, it should be possible to test these and use the composite result for reporting reading proficiency. (Weir 2005:88)

Background

In his much-cited 2005 book, Cyril Weir took a fresh look at the principles underlying language test validation. One part of this review entailed redefining the broad categories conventionally used in discussions of language test validity. The notion of 'reliability' now, in Weir's socio-cognitive framework, became *scoring validity*. Further criteria included the *consequences* of a test (e.g. the use that might be made of the results and the washback effects upon instruction) and the need to trace *comparisons* a) between different versions of the same test and b) between a given test and others targeting the same skills, domain and population.

Most relevantly to the present chapter, the framework marked out a distinction between two aspects of construct validity:

- the characteristics of a test: classed as *context validity* and embracing aspects of the input and the task, but also the circumstances under which the test took place
- the perceived goals of a test: classed as *theory-based validity* and embracing not simply the linguistic content targeted but also the behaviour that the test aimed to elicit.

Though he characterised 'theory-based validity' in terms of behavioural as well as linguistic criteria, Cyril Weir did not initially make a connection with a growing body of research, particularly in the USA, that concerned itself with *cognitive validity* (a term introduced by Glaser in 1991). Behind these studies was a relatively straightforward question which had previously been little addressed in educational testing. Does a given test simply measure knowledge of pieces of information or does it also measure the candidate's command of the cognitive processes associated with employing that knowledge? To put it more concretely, to what extent does a test of Medicine or Physics simply tap into factual information associated with those subjects? To what extent

can it be claimed that the test taker is also required to demonstrate that they are capable of thinking like a doctor or a physicist? The applications to the testing of second language skills will be evident. Indeed, the issue represents familiar terrain in the applied linguistics literature, where Anderson's (1983) distinction between declarative and procedural knowledge has been much cited and commentators have frequently contrasted, on the one hand, explicit knowledge of a language and, on the other, implicit knowledge and language use (Ellis 1994:1–31, Ellis et al 2009, Rebuschat (Ed) 2015).

Once Cyril Weir's attention was drawn to the relevance of this parallel field of enquiry, he was quick to incorporate it into his own model – later replacing the term 'theory-based validity' with 'cognitive validity'. A willingness to make interdisciplinary connections of this kind was a hallmark of his academic thinking. It was accompanied by a degree of rigour when taking on unfamiliar ideas that is perhaps not as common as it should be in our field. Cyril was always aware that the knowledge of any individual academic commentator is necessarily finite; and that, in extending one's net, one needs to be willing to draw on and take due account of the views and expertise of specialists in other domains. This was the origin of the present writer's participation in a series of four validation exercises relating to the tests of the Cambridge English suite (Shaw and Weir 2007, Khalifa and Weir 2009, Field 2011, 2013). Their aim was to build bridges between cognitive evidence concerning the nature of skilled language behaviour and current approaches to the assessment of L2 language skills.

There was another reason for Cyril Weir's growing interest in cognitive criteria. His 2005 proposal for new validation categories was accompanied by some important insights into how future validation exercises should be conducted. He was very sensitive to the dangers of relying heavily on conventional score-based statistics to determine the construct validity of a test. A spread of scores might indeed indicate that a test discriminates between candidates; but there can be no guarantee that the differences demonstrated relate wholly or even partly to the L2 skill being targeted. The challenges that separate out the performance of test takers might equally well derive from aspects of the item, format, rubric or text that are not directly relevant to the construct.

From this, Cyril drew two major conclusions, both closely linked to behavioural criteria. Firstly (2005:Chapter 3), he stressed the importance of approaching test design from the outset with a clear understanding of the construct being targeted, one that was ideally based upon sound empirical evidence. This *ab initio* approach could lend rigour to any validation claims that the test designer might make.

Secondly, he argued (2005:Chapter 4) for a new *post hoc* approach to validation that did not simply rely on scoring data but that matched test content and test results against a well-attested model of the skill being targeted. This latter approach has proved especially fruitful. As noted, it has

informed a number of validation studies of tests in the Cambridge English suite. From these exercises, methodological approaches and analytical criteria have emerged which seem set to influence future thinking on language test validation.

The present chapter first provides a wider background to the notion of cognitive validity and describes its recent role in tests of second language skills. It then gives an account of behavioural models of the four language skills which have been employed in recent validation exercises. As already indicated, such models can be used both to guide the initial design of a language skills test and to provide a yardstick against which test content can later be systematically measured. An early example of this type of application is to be found in Weir's exploration of the cognitive processes underlying the writing skill, which will be briefly reviewed. But particular attention is reserved for reading, his principal area of interest, with comments on the way in which notions put forward by him have come to inform thinking. The chapter then goes on to exemplify some of the methodological approaches to cognitive validation that have emerged over several years of research and discussion. A final section reflects on the future roles of verbal report and computer technology in shedding light on learner cognitive behaviour under test conditions.

The chapter as a whole attempts to represent an important area of Cyril Weir's legacy to second language testing; it is also a testament to a fertile and memorable partnership between a language testing specialist and a psycholinguist.

Cognitive validity

The concept of cognitive validity (Glaser 1991) came to the fore in the 1990s. It extended conventional notions of construct validity to take account of the extent to which a test requires a candidate to engage in cognitive processes that resemble those that would be employed in non-test circumstances. As already noted, there are clear applications here to educational contexts: for example, does an achievement test in a history course demonstrate that its students have come to think like historians or simply that they can report what happened?

The issue has had a considerable impact upon approaches to educational measurement in the USA. As well as helping to define behaviour relevant to success in specific subject areas, it has been applied to tests of scientific and logical reasoning (Linn, Baker and Dunbar 1991, Ruiz-Primo and Shavelson 1996, Baxter and Glaser 1998, Thelk and Hoole 2006). It has also been used to investigate the learning process: including aptitude tests (Snow, Corno and Jackson 1996), mastery of lesson content (Koskey, Karabenick, Woolley, Bonney and Dever 2010) and the ability to map from input to concept

(Ruiz-Primo, Schultz, Li and Shavelson 2001). In professional contexts, it has shed light on the cognitive processes that underlie tests of clinical diagnosis (Gagnon et al 2006) and decision making (Larichev 1992).

Field (2013:79) describes the relevance of cognitive validity to a language assessment context, particularly where a test's aim is to predict future behaviour:

> Cognitive validity is a particular concern in the case of tests whose scores are employed predictively to indicate the test taker's suitability for a future university place, for a job in a domain such as business, medicine, teaching or tourism or for acceptance under an immigration programme. It is not enough for such tests to demonstrate that a test taker has reached a criterial level of language knowledge; they must also be capable of demonstrating that the test taker is capable of linguistic behaviour that meets the requirements of the target context. If there is a significant qualitative difference between the processes elicited by the test and those demanded by the context, then the ability of the test to predict performance is open to question.

It is important to distinguish this concept from ecological validity, which entails a more general concern with the extent to which a test can be seen to replicate real-world conditions. What is at issue in cognitive validation is whether the tasks proposed by the test designer *engage mental processes that are adequately representative of those that a language user would employ in a real-world context*.

Within language testing, it is also important to distance the notion of cognitive validity from certain statistically based approaches to construct validation (e.g. factor analysis, multiple regression and structural equation modelling). These might entail identifying a number of test features that are believed to contribute to difficulty and then deriving statistical evidence of the extent to which each contributes to the scores obtained. This type of analysis can provide useful guidance to test designers as to the features which can be most easily manipulated in order to increase the difficulty of a test. That said, research results indicating which factors carry the most weight tend to vary across studies – partly, no doubt, due to variations in the number and range of factors investigated by different research teams. For an account in relation to reading, see Khalifa and Weir (2009:35–37), who attribute divergences between researchers to wider design features such as the population sampling, the methodology used in data analysis and the particular tasks involved. In relation to the cognitive demands of a test, one might add that hard and fast conclusions cannot be drawn from this type of data unless the effects upon scoring of test characteristics (vocabulary, length, format, item content etc.) are linked to specific cognitive processes to which they give rise.

A second qualification is that, with its use of items, scores etc., this type of analysis is closely associated with a test-taking context. It thus makes no claims in relation to the main concern of cognitive validation: the extent to which within-test behaviour can be said to be representative of real-world performance. Khalifa and Weir make the point very tellingly (2009:37): 'The concern in this psychometrically driven approach is thus not with the actual components of the reading process that are necessary for comprehension but with the factors which can be shown statistically to contribute to successful performance in the specific tests of reading under review . . . Thus the data examined is a measure not of successful reading per se but of successful performance in the test. The factors underlying the latter do not necessarily hold true for reading activities that take place in the real world.'

They go on to raise a further point that was long a particular interest of Cyril Weir's (see e.g. Urquhart and Weir 1998). A psychometric approach cannot represent the wider cognitive and metacognitive decisions that competent readers make in choosing a *reading style* that is appropriate to the current task. Should they opt for an expeditious style or a careful one? Does a specific item require a local approach to reading or a global one? It is relevant to ask whether the tasks used in a test tend to focus heavily upon a single style (e.g. local and careful) or demand a variety of reading decisions.

If an approach such as multiple regression cannot provide unambiguous insights into the mind of the test taker, where should one turn instead for information about the components of a language skill that a test is or should be targeting? An obvious solution a generation ago might have been found in the lists of *sub-skills* proposed by a number of sources, which had considerable impact upon the way in which skills were handled in the classroom. They began with Munby (1978) and included for reading, Grellet (1981) and Nuttall (1996) and for listening, Richards (1983). These taxonomies played an important part in drawing the attention of both teachers and testers to the importance of language *performance* as against language knowledge; but their drawback is that they were based not upon hard evidence but upon the intuitions of commentators, albeit informed and experienced ones. They thus do not provide an empirically supported framework against which test performance can be measured in any systematic way.

When the notion of cognitive validity was applied to L2 testing, it became apparent that wider fields of information needed to be drawn upon than those associated with conventional notions of construct validity. Prominent among them were several areas of psycholinguistics – including speech science, information processing, working memory, child language development and theories of how L1 speakers acquire the written skills. Areas more familiar to applied linguists included phonetics (both productive and receptive) and discourse analysis.

Psychology and language skills

Models of expert performance

The value of drawing upon psycholinguistic sources in particular lies in the fact that, over many years, researchers have constructed detailed cognitive models for all four language skills. To be sure, there remain some grey areas on which they fail to agree. But in general these models are not just based upon well-supported theoretical constructs; they are also underpinned by detailed evidence, collected through painstaking studies of small-scale contributory processes. In other words, the models have strong empirical support.

Such models provide coherent accounts of the way in which a competent reader, listener, speaker or writer comprehends or produces language. They can thus be said to provide a detailed profile of the end-point towards which a second language learner is aiming: namely, a high degree of fluency in employing the processes that make up the skill.

Most importantly, they provide a systematic framework for validating current L2 tests in terms of how accurately they represent or elicit certain aspects of real-world language performance. Two clearly defined lines of enquiry have emerged:

> a. *Similarity of processing*. Are the processes adopted during a test sufficiently similar to those which would be employed in the target context? Or do candidates adopt additional processes that are a product of facets of the test (procedure, test method, item) rather than part of the normal operations associated with the construct being tested? . . . In other words, is there what Messick (1989) terms *construct irrelevance*?
> b. *Comprehensiveness*. Do the items in the test elicit only a small sub-set of the cognitive processes that a language user would employ in a natural context? Or do they tap into a sufficiently broad range of such processes for the test to be deemed representative of real-world behaviour? This might be a reflection of the method employed; but it might equally be a question of how diverse the test items are . . . This would seem to relate to Messick's (1989) concern over possible construct *under-representation*. (Field 2013:80)

In addressing these questions, cognitive models of language use make an important contribution to Weir's (2005) call for informed *post hoc* validation. But they can also serve an *ab initio* role in influencing the design of future language tests, and ensuring that they target a range of processes for which there is reliable real-world evidence.

A framework for language acquisition

The approaches described so far relate to language behaviour. However, a cognitively based approach to test design and validation can also shed light on issues of acquisition. There is an interesting body of theory in general cognitive psychology which aims to account for what is termed *expertise* (Anderson 2000:256–288, Ericsson and Smith (Eds) 1991, Feltovich, Prietula and Ericsson 2006): explaining, for example, how a novice driver manages to convert controlled processes demanding attention into automatic ones that do not. The value of automatic behaviour is that it is rapid and accurate, that it requires minimal concentration and that it makes few demands upon an individual's working memory.

Treating language as a form of expertise, psycholinguists have applied the same principles to describing how those acquiring a first or second language achieve a basic level of fluency in the four skills. Broadly, the argument (Segalowitz 2013) is that early language is constrained by the learner's need to focus effortful attention on smaller units and on the process of assembling those units when delivering them as a speaker/writer or decoding them as a listener/reader. This limits the capacity of the individual to lend attention to wider patterns of thought relating (e.g.) to context and line of argument. Evidence supporting this kind of developmental path has long been available from research into the L1 mastery of the learned skills of writing and reading (see e.g. Bereiter and Scardamalia 1987, Oakhill and Garnham 1988). Khalifa and Weir (2009:48) apply it to the L2 reading experience: 'The attentional resources of a reader are finite and, at least in the early stages of L2 development, one might expect a large part of those resources to be diverted towards low-level considerations concerning the linguistic code'. Evidence (Field 2019:16–18) illustrates that second language listening is shaped by the same constraint.

An understanding of the processes that drive skills acquisition can thus provide test designers with a systematic view of what a language learner is likely to be capable of at different levels of proficiency. This information can usefully supplement the types of descriptor we already possess (e.g. those of the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001) and of various test boards), which are the outcome of the reports of highly experienced commentators but are not underpinned by an established theory of what lies behind these patterns of development. By drawing upon a model of a skill like those described above, a cognitively informed approach enables us to clarify which sub-components of the skill we can reasonably expect a learner to have mastered at a given proficiency level. Broadly, what this means is that test items at lower levels should chiefly entail the use of *perceptual* processes rather than *conceptual* ones. To give the example of reading, the items should require test takers to obtain mainly

factual information based on making literal sense of a text at word, clause or sentence level; but should not require them to make inferences or to report a wider line of argument.

In effect, this adds an additional (third) strand of enquiry to a cognitive validity exercise and does so in ways that assist both of Cyril Weir's 2005 lines of attack. *Ab initio*, it provides additional descriptors that can sensitise a test developer to the types of text and task that learners are cognitively capable of handling at different proficiency levels. *Post hoc*, it enables us to add a further validation question to the two cited above, namely:

c. Are the cognitive demands imposed upon test takers at each proficiency level *appropriately calibrated,* in relation to the performance features that might be expected of a second language user at that level? (Field 2013:80)

Cognitive validity in practice

Recent investigations into the cognitive validity of tests of the four L2 skills have thus illustrated a need for:

- clear and empirically supported models of each skill, which provide frameworks that represent how an expert language user performs. Against them, one can measure, in relation to any given test of L2 skills: a) the stated and implicit presuppositions that underlie the test; b) the test content; c) the task types and items employed; d) expectations as to L2 performance at different levels; and e) test taker performance observed or reported.
- a principled way of accounting for how L2 performance at different proficiency levels can be expected to diverge from the expert template.

It is not just a matter of establishing how effective a particular task is at emulating the type of behaviour which a real-world speaker, listener, reader or writer would employ. It is also important to match the behaviour elicited by the task against a clear understanding of what one can reasonably expect of a learner at a given stage of development. One might also note the need, with specific populations, to understand the nature of the discourse that they will have to produce and comprehend or the limitations that they may bring to the test. In academic and professional contexts, this might entail evidence of the domain-specific language behaviour in which a candidate will typically have to engage. In young learner contexts, it might entail evidence of the likely cognitive development of candidates, especially those aged 8 to 12 (Field 2018a, 2018b).

Cognitive models of language processing

Background

An important aspect of recent test validation exercises influenced by Weir's socio-cognitive framework has thus been the building of bridges between psycholinguistic theory and the criteria that guide the testing of the four skills. Established cognitive models of the four language skills provide an account of the behaviour towards which instructors hope they are leading L2 learners. They also provide a framework for examining precisely which processes a given test taps into and which may have been overlooked or misrepresented.

Much information relating to the acquisition of the two skills that are *taught* in L1 (reading and writing) is available from accessible L1 educational and psychological sources such as: Snowling and Hulme (Eds) (2005), Rayner, Pollatsek, Ashby and Clifton (2012:277–343), MacArthur, Graham and Fitzgerald (Eds) (2006). That said, an L2 commentator may have to factor in some additional considerations:

- the influence of skills acquired and used in a learner's L1 may delay or impede the acquisition of similar skills in L2 (particularly motor skills and word recognition/spelling)
- in a sub-set of early-stage candidates, allowance may have to be made for lack of familiarity with the alphabetic system and/or with Roman script
- in tests of young learners, attention may need to be given to candidates' current level of achievement in L1 reading and writing.

Areas of research that may be less familiar relate to the skills that are *acquired* in L1. In point of fact, there has been quite intensive cognitive and phonetic research since the mid-1960s into both L1 speech production and L1 speech perception. Speech scientists have reached a high level of agreement as to the fundamental processes in each; the problem has been that these findings tend to be presented in a rather terminologically dense way and in journals that may not fall within the interest areas associated with applied linguistics.

Cognitive models of the language skills generally follow a basic principle of *information processing*. This entails a view of both language perception and language production as a set of operations that gradually transform pieces of information from one form into another. In terms of reading, the initial form would be a string of letters on a page or screen and the final form would be a concept (or a set of concepts) in the mind of the reader. Reversing the process, in writing the language user would begin with a concept or set of concepts and the outcome is a string of letters on the page or the screen. The processes in question are often presented in the form of flow charts. Figures 1 and 2 represent, in simplified form, the general models which underpinned

Figure 1 Simplified model of language reception (adapted from Field 2015)

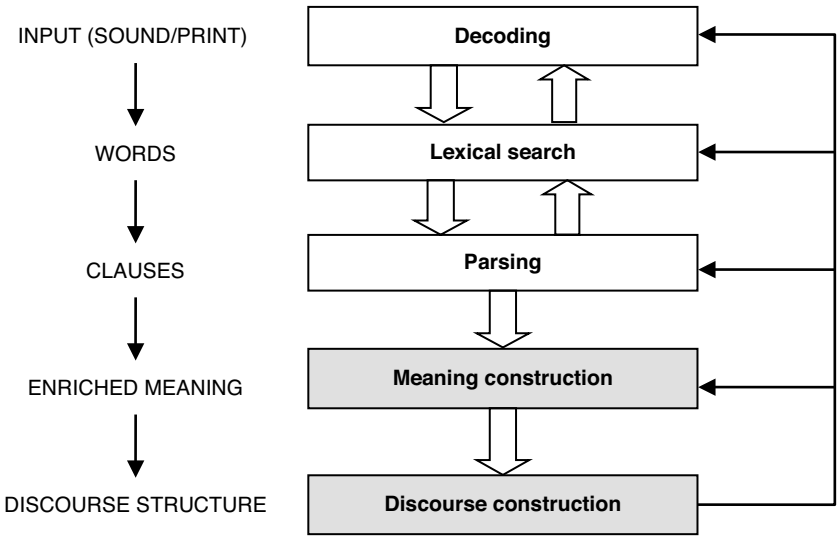
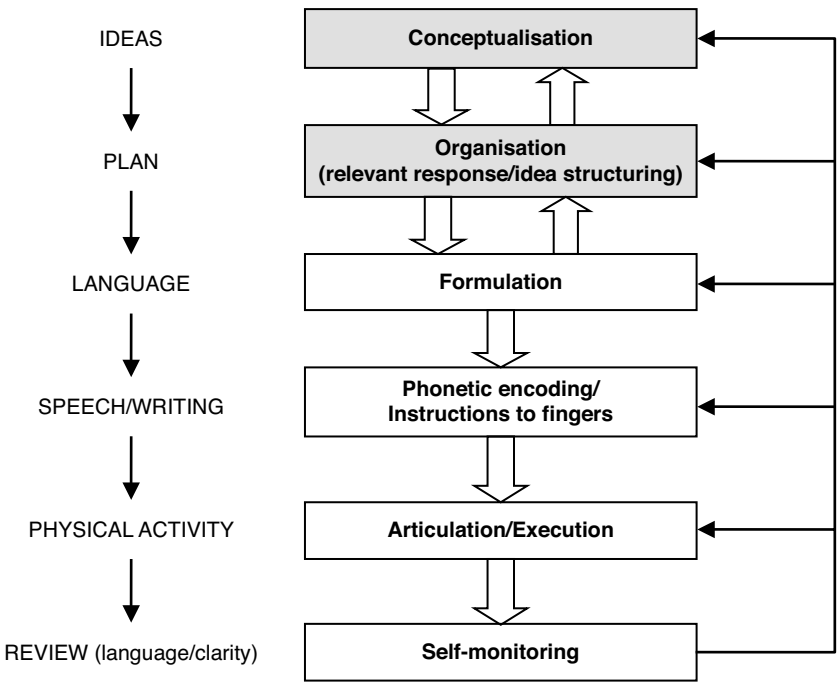


Figure 2 Simplified model of language production (adapted from Field 2015)



the behaviour described in the course of a series of validation exercises in the Cambridge English SiLT series (Shaw and Weir 2007, Khalifa and Weir 2009, Field, 2011, 2013). A few of the terms have been changed for the purposes of this account and in the interests of clarity for the reader. Grey tones indicate that the information being handled is in the form of ideas rather than language.

At this point, an important caveat needs to be added. The steps through which (say) a reader develops the raw information on the page are often referred to as stages or phases in the reading process. This has sometimes led to the assumption that the process is unidirectional:

> letters → recognition of a word's form → word meaning → clause meaning → interpretation of clause → inserting the clause into a wider pattern of discourse.

In fact, current cognitive models of the four skills are interactive and acknowledge that information can flow in both directions. For example, establishing meaning at clause level may indicate to a reader that they have misread a word (or indeed might help an L2 reader to work out the meaning of an unfamiliar word). Similarly, when the overall line of argument seems to conflict with a clause that has just been read, a reader may decide to look back to check that the clause has been correctly understood. This kind of effect is represented by the upward arrows in the figures.

Our understanding of how processing operates is not helped by the loose use of the terms 'bottom-up' and 'top down' in the L2 literature (Field 1999). Strictly speaking, these terms refer to the directions of processing just mentioned; but they are sometimes used confusingly to make a distinction between *lower-level* or *perceptual processes* (in reading, recognising words, identifying a grammar pattern, etc.) and *higher-level* or *conceptual processes* (such as interpreting the writer's intentions). The more precise terms are adhered to here.

A cognitive account of writing (Shaw and Weir 2007)

Weir's first exploration of the links between cognitive theory and second language testing focused on writing. The account provided by Shaw and Weir (2007) drew principally upon a model of the skill proposed by the psycholinguist Kellogg (1994, 1996). However, it had recently been suggested (Field 2004) that the Kellogg model was in certain respects too closely aligned to Levelt's much-quoted model of speaking (1989). As a result, it tended to ignore two features which sharply distinguish writing from speaking: the fact that it is not time-constrained as speaking is, and the fact that it entails decisions at a text-planning level as well as at the level of the individual paragraph

or sentence. Always alert to fresh ideas, Cyril amplified the first two stages shown in Figure 2 into:

- *macro-planning*: generating ideas and identifying constraints (genre, readership, goals)
- *organisation*: sequencing the ideas; identifying relationships between them; determining their relative importance
- *micro-planning*: broadly, planning at paragraph and sentence level, but also with constant reference to the writer's goals.

On this basis, Shaw and Weir focused on a set of six cognitive processes:

> *Macro-planning – Organisation – Micro-planning –Translation*[1] *– Monitoring – Revising*

They intentionally omitted three important operations of writing (storing a planned clause in the mind, turning the plan into instructions to fingers and executing the plan). These operations were deemed to be so automatic as to be inaccessible to report and, what was more, incapable of being matched against any specific features of test content or task, as part of a validity argument. In taking this step, Cyril presciently anticipated what has been a major obstacle to associating certain aspects of the productive skills with test features: namely, the fact that some of the operations employed in those skills are so highly integrated or (in the case of speaking) so time-constrained that it is impossible to separate them out or to align them with any specific feature of test design. Some commentators identify only three in writing: planning–translating–reviewing (Torrance and Jeffrey 1999:5–7).

Drawing upon L1 sources, Shaw and Weir also attempt to identify aspects of learner behaviour that might indicate progress in L2 writing proficiency. From Hayes and Flower (1986), they report evidence that skilled writers give more attention to coherence and argumentation during planning and spend more time self-monitoring and revising than do less skilled. They also cite an influential account by Scardamalia and Bereiter (1987) of a shift in L1 writers between an early period of simply reporting knowledge and a later one of structuring it. Shaw and Weir are quick to associate phenomena such as this with working memory and automaticity: suggesting (2007:41) that, at least in the early stages of L2 development, one should expect a large part of the limited resources of a writer to be diverted away from planning and monitoring towards lower-level linguistic considerations.

1 With hindsight, the retention of Kellogg's term *translation* to refer to the process of encoding a message into language was unfortunate and potentially misleading in an L2 context.

A cognitive account of reading (Khalifa and Weir 2009)

Let us now look more closely at the cognitive framework to which Cyril Weir gave most thought: his model of reading, which has influenced a number of recent research studies (e.g. Bax 2013, McCray and Brunfaut 2016). This will also serve to outline the contribution which he made to extending the framework provided by psycholinguists.

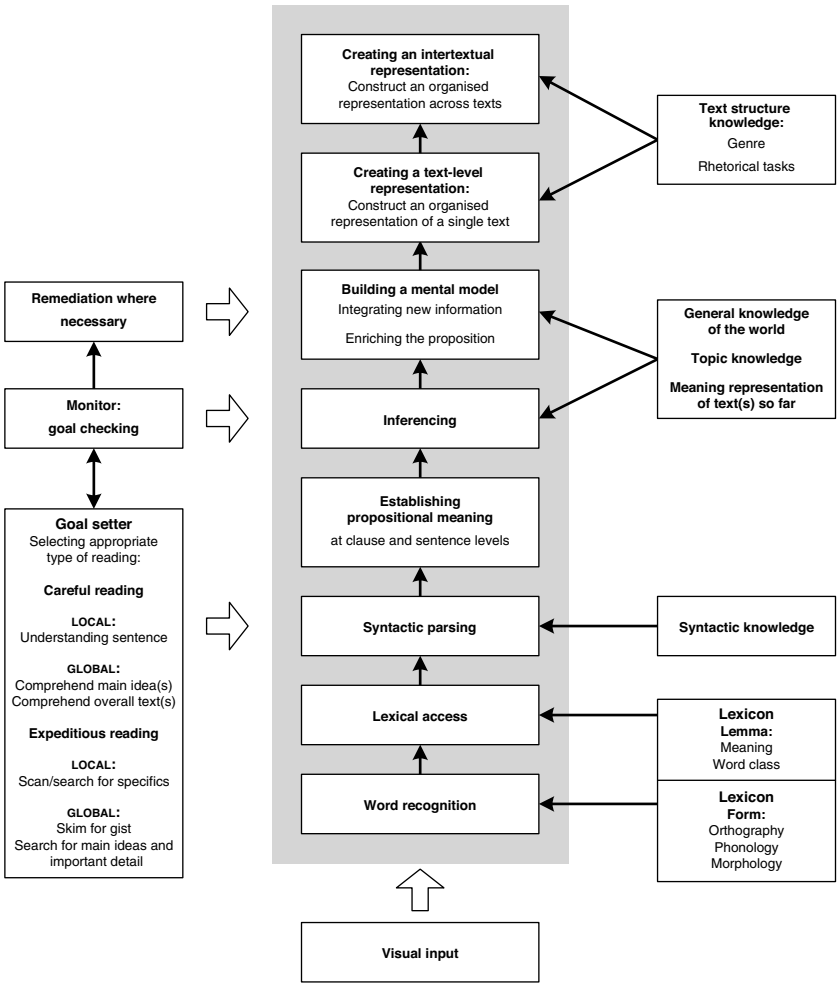
Figure 3 is taken from Khalifa and Weir (2009:43) and also appears in a posthumous publication (Weir and Chan 2019). At first sight, the model might seem bewilderingly complex; but in fact its central core is the highlighted middle column, which represents the five principal operations that contribute to competent reading. They are:

- *Decoding*: Matching the sight of a word on the page against a representation of a word's written form that is stored in memory.
- *Lexical search*: Opening up a cache of information about the word – word class, morphology, possible collocates, and above all the word's sense or range of senses.
- *Parsing*: Imposing a recognisable syntactic pattern (an easy example would be SVO) on a meaningful group of words – a phrase, a clause or a sentence.
- *Meaning construction/building a mental model*: Turning a basic proposition into a contextualised piece of information.
- *Discourse construction/building a text-level representation*: Embedding a new piece of information in a wider pattern of argument relating to the text as a whole.

Weir made some changes and additions to the standard set of operations; and it is of interest to note what he contributed. He chose to divide the operation of meaning construction into two distinct processes: firstly, 'inference' (i.e. supplying information not explicitly expressed by the writer) and secondly, 'building a mental model' by means of world knowledge and awareness of the current topic. Here, he seems to have felt the need to distinguish two distinct types of cognitive behaviour within the larger operation.

More importantly, he added an additional category to the five conventional phases, in the form of an operation largely overlooked by psycholinguists: namely, *creating an intertextual representation*. This relates to a reading goal that is critical in academic contexts but also more important than generally realised in other domains (one thinks of comparing project proposals, political manifestos, hotel facilities, CVs, product reviews, etc.). It is a type of reading whose relevance has only recently been recognised

Figure 3 Weir's cognitive model of reading (Khalifa and Weir 2009)



(see e.g. Perfetti, Rouet and Britt 1999, Lacroix 1999, Strømsø and Bråten 2002, Goldman 2004) and one that future testers of L2 reading (especially in English for Academic Purposes (EAP) contexts) do well to take into account. Indeed, it has already informed the development of reading-into-writing tasks which require test takers to draw upon multiple sources (Hartman 2005, Chan 2018). Khalifa and Weir (2009:53–54) make the telling point that this kind of intertextual operation imposes additional cognitive demands upon the reader. The line of argument and the conceptual links that support the building of a single-text representation have to be considerably supplemented

when a reader goes on to trace connections between texts and when information has to be selected, evaluated and redistributed.

Besides adding to the established information-processing profile, Cyril, ever a lover of detail, added two major features with a view to making his framework as informative as possible to those working within language testing. These are represented in the columns that appear to the left and right of the flow chart. The one on the right indicates the sources of information on which readers draw during the various phases of processing. It draws attention to a distinction between the part played by reader *behaviour* as displayed in the central column and that played by reader *knowledge*. This distinction sometimes becomes blurred in linguistic accounts of the language skills.

Psycholinguists conducting research into the nature of skilled L1 reading have tended to focus on the underlying cognitive processes; and the part played by *metacognition* has been rather downplayed (exceptions include Cain and Oakhill (Eds) 2008, Westby 2004). This is curious because reading is not time-constrained in the way that listening is, with the result that metacognitive decisions and choices play an important part in reader behaviour. The left-hand column of Weir's model serves to correct this oversight by foregrounding two types of metacognitive activity. Firstly, it highlights an issue that (as already noted) was a long-term interest dating back to Urquhart and Weir (1998): the types of *reading style* in which an adept reader is likely to engage. Those styles might represent decisions to read locally for detail or globally for main points and line of argument. They might similarly represent a reader's intention to read rapidly and shallowly for general information or in depth to be sure that information has been fully mastered. Note the terms 'decision' and 'intention': unlike the processes in the main column, these are behavioural directives that the reader may be able to report and is certainly able to control.

Reading style underpinned many of the skills-based reading tasks that were employed as part of the movement towards skills-based instruction in the 1990s (see e.g. Grellet 1981, Nuttall 1996). Learners were routinely encouraged to approach tests strategically: skimming them for gist, then scanning them for key words or items of information and finally reading them in depth. This was said to be a process of test familiarisation that would assist learners in real-world contexts – indeed the techniques were derived from first-language study skills materials. Yet this angle upon reading has had curiously little impact upon thinking in testing circles. What, in effect, Weir signalled in adding reading style to his model was that the ability to handle a text flexibly and with relevance to the reader's goals might well be an important strand of overall reading competence. A second, and equally important, message was that testers should look critically at the items they devised, to ensure that they did not unduly favour a single type of reading. Given the formats used in comprehension testing, one might expect a heavy bias in favour of local in-depth reading.

The other feature which Weir introduced into his account of reading is also metacognitive. Psychological models of writing and speaking conventionally assume a role for self-monitoring (see Figures 1 and 2). In writing, an important part of the exercise entails revisiting and editing a draft (Bereiter and Scardamalia 1987:278–293). Similarly, speakers self-monitor throughout their turn to ensure that the utterance is a) accurately formed in relation to what was planned, and b) comprehensible to the listener (Levelt 1989, 1999). The role of self-monitoring in reading seems relatively under-explored (though see Yuill and Oakhill 1991:Chapter 6). Here again, account needs to be taken of the fact that reading is not time-constrained like listening and allows the reader ample opportunity to look back to check understanding. It is difficult to imagine how this aspect of the skill might be targeted by a task or item; but it is surely a consideration when determining how much time should be allocated to a given reading task.

Cognitive validation: an emerging methodology

Basic principles

We now consider how the cognitive principles outlined so far have enabled investigators to add a new strand to test validation – one that focuses on the extent to which a test elicits processes representative of those of the world beyond the classroom.

As mentioned earlier, cognitive validation exercises involving the tests of the Cambridge English suite led to the conclusion that any such task needed to address three questions:

a. How *representative* is the language behaviour elicited by the test in terms of what the test taker would need to deploy in real-world conditions?
b. How *comprehensive* is the content of the test in replicating the range of behaviour associated with competent performance in a given language skill?
c. How *well calibrated* across proficiency levels is the language behaviour elicited by a test?

It might also be necessary, with certain populations, to take account of the nature of the discourse that they have to produce/comprehend or the limitations that they might bring to the test. An example of the first would be the domain-specific language behaviour in which a candidate typically has to engage in an academic or professional context (Field 2019:99–104). An example of the second might be an understanding of the likely cognitive and literacy development of young learner candidates, especially those aged 8 to 12 (Field 2018a, 2018b).

The virtue of using an empirically supported model of one of the language skills as a point of departure is that it enables one to systematically match a set of contributory operations (decoding, lexical search, parsing etc.) against available evidence of the characteristics of a test. Possible sources for that evidence include:

- specifications as to expected L2 performance at different levels (including published descriptors, claims about CEFR compatibility and raters' guidelines)
- the presuppositions that underlie the test, as evidenced by the test board's handbooks, publicity and instructions to item writers
- test content (input texts for the receptive skills; stimuli for the productive ones)
- the task types employed and instructions to item writers regarding them
- the items featured and the level of response they elicit
- the aspects of performance targeted in raters' guidelines for the productive skills
- observed and reported test taker performance.

It will be noted that features specifically associated by Weir (2005) with *context* validity are included in this programme of cognitive analysis. Early validation studies of the writing and reading tests of the Cambridge English suite (Shaw and Weir 2007, Khalifa and Weir 2009) adhered quite rigorously to the distinction between 'context' and 'cognitive'. However, time and experience have demonstrated a much more permeable relationship between the two areas. Several aspects that one might normally class as 'context' have important cognitive implications that cannot be ignored. To give a few random examples:

- Vocabulary and indeed grammar are not just about what a candidate can be taken to 'know'; but also about what they are able to *recognise* when hearing connected speech and *retrieve* under time pressures when producing it.
- Because a listener or reader has to parse a text word-by-word as it is heard, longer utterances/clauses increase processing demands.
- Conventional test formats sometimes engage cognitive processes that are specific to a testing context and have no real-world correlates.
- Certain types of discourse are easier to process in tests of reading and listening because of more transparent conceptual links between units of information. These same discourse types are also easier to assemble in tests of speaking and writing.
- The information density and degree of subordination in a reading or listening passage determine how demanding it is for a test taker to extract meaning.

- Features of the recording (including number of voices, speech rate and accent) can have a major impact upon listening difficulty.

A few sample approaches will now be described that directly address the three major questions just cited. They exemplify some of the issues of methodology that have come to the fore in the course of cognitive validation projects.

Representative?

A cognitive analysis needs to take account of the *language content* of a test – matching it not against an aspirational grammar or vocabulary syllabus but against whether it resembles the language of the real world in which candidates will have to perform. In tests of listening, this generally raises issues of comparability rather than of strict authenticity. Where scripts are written by item writers, they often lack the discourse features of natural speech (see Gilmore's 2015 findings on language teaching materials). Elsewhere, test boards might transcribe authentic texts and re-record them in a studio to sidestep the problem of obtaining permissions. In both cases, a major criterion for any validation exercise becomes to what extent an actor operating in studio conditions with a written script in their hand has succeeded in capturing features of connected speech such as planning pauses, intonation patterns, weak forms, elisions and reductions. This line of enquiry requires a validator to listen with some care to the *recording* as the form of transmission that the candidate encounters, rather than relying on a script.

In addition, item writers working on listening and reading tests quite frequently revise semi-authentic scripts of the kind described – partly to control for unfamiliar vocabulary and language, but often to insert multiple-choice distractors. A second area of concern is therefore whether this last adjustment increases the information density of a test to the point where idea units are more densely packed than normal and the candidate is likely to find them hard to retain.

Questions can also be raised about conventional *task formats.* Tests of reading and listening tend to rely on familiar types: multiple choice, gap fill, true/false, multiple matching. A cognitive perspective serves to shed light on the processes elicited by these familiar formats and to illustrate how closely they do or do not resemble real-world communication.

Consider multiple choice. The test taker has to engage in the highly artificial process of committing three or four propositions to mind, noting how they differ and seeking a paraphrase for one of them in the wording of the text. As a cautious (or curious) human being, they are also oriented towards trying to discredit the incorrect ones. A similar concern attaches to gap filling. It may appear to emulate the note-taking of an academic learner, but in this case the notes are not the learner's own. It represents a classic example of what psychologists term a *divided attention* situation (Pashler and Johnston

1998). A reader is encouraged to track back and forth between text and item; while listeners (Field 2012) have to engage three skills simultaneously and to do so under pressures of time. Gap filling also effectively provides the candidate with an advance summary of what will be encountered in a text – and does so in the case of listening by means of a written medium that is usually easier to process than a spoken one.

There is no easy solution. Major test boards are committed to these formats, which, to be sure, are familiar to candidates and easy to mark. But it is important, at the very least, to acknowledge the value of using a *variety* of formats. Each demands its own set of extraneous cognitive processes which may accord with the mental set of some candidates but not necessarily all. One can also argue a strong case for avoiding these formats in local testing, where the need for ease of delivery and ease of marking is not present. Local instructors designing progress tests do well to avoid them and rely instead on open-ended questions, transcription, note-taking, etc.

The formats used to test the productive skills are not quite so problematic. Even so, tests of speaking sometimes rely disproportionately on a single format (Q&A with a stranger), which replicates the circumstances of one relatively rare type of speech event – an interview. Where there are monologue speaking tasks, fine decisions also have to be made about preparation time. The time allocated has to be long enough to allow the speaker to *conceptualise* and *plan*; but not long enough to rehearse utterances that can be produced verbatim. Writing is probably the only one of the skills where the task conditions approximate to those of a real-world activity; here, the major difference lies in timing constraints.

Comprehensive?

The models shown in Figures 1 and 2 provide an outline of the major operations entailed in language reception and production. An obvious way of establishing how comprehensively a language test represents one of the skills is to match these operations against those reported by test takers, in order to establish how many are actually present.

However, this is not as simple as is sometimes assumed. Within the receptive skills, the perceptual operations (decoding, lexical search, parsing) are intrinsic. A reader simply cannot make sense of a text – or a listener of a speech signal – without engaging these processes. So it is meaningless to try to establish whether a reader/listener has or has not used them. They assuredly have: the skill cannot function without them. There have been some misconceptions on this score, where studies (e.g. Brunfaut and McCray 2015:36–38, Holzknecht and Eberharter 2017:24–27) have used learners' verbal reports to seek evidence of processes that are an inextricable part of the skill under discussion.

Leaving aside concerns about cognitive validity, a major advantage of the

conventional formats used in testing reading and listening is that they rely upon a set of questions. Each question can be thought of as focusing on a key in the text in the form of a word, a group of words or a line of argument. An approach adopted in recent cognitive validation exercises (Field 2013, 2018a, 2018b) has therefore been to examine items across sample tests in order to specify what level of processing each taps into. Relatively uncontroversial judgements can be made on the basis of a set of concrete questions. Taking listening as an example, they might be:

- Does this item require the candidate to distinguish two phonologically similar words? (*fifteen* vs *fifty – nine* vs *five – hungry* vs *angry*) [DECODING]
- Can this item be answered by information at the level of a word or formulaic phrase? (*drive*/*fly*/*walk – agree*/*can't agree – next to*/*in front of*) [LEXICAL SEARCH]
- Can this item be answered by factual information at the level of an utterance/a clause? (*I won the match – I decided not to go – It wasn't a very good film*) [PARSING]
- Can this item only be answered by understanding functional language or by interpreting language, adding to it or placing it in context? (*Would you mind if . . .? – He's a Sunday driver – That's not what I meant – I'm not sure that's a good idea – I'll think it over*) [MEANING CONSTRUCTION]
- Can this item only be answered by making connections between two conjoined pieces of information or two widely separated pieces, by tracing a main point or line of argument, by reporting a point of view or a speaker's general attitude? [DISCOURSE CONSTRUCTION]

A similar set of item targets can be used for reading – though with the proviso that 'decoding' is likely to be relevant at only the most basic proficiency levels and in cases where the candidate's L1 employs a different alphabet or writing system.

Of course, the same approach cannot be extended to tests of the productive skills, where the tasks are much broader and not itemised. What is more, the productive skills are even more difficult to separate into their component operations. In both speaking and writing, the stages of production (formulation/encoding/signals to articulators or fingers) are very tightly integrated. In the case of speaking, they also take place under extreme pressures of time.

This entails a more interpretive approach. Experience suggests that several sources of evidence can be invoked when considering how comprehensively a given test represents relevant operations. They include: sample tests, instructions and prompts provided to candidates, rating scales, instructions to item writers, test specifications in handbooks and the range of topics and discourse types covered. One means of approaching rating scales in particular is to trace correlations between internal cognitive processes and some of the

external evidence employed when a learner's competence is being judged. In speaking, for example, one might find *formulation* represented in terms of 'accuracy', 'comprehensibility', 'functional language' or 'language chunks'. *Phonetic encoding* is clearly covered by references to 'pronunciation', 'intonation', 'rhythm' and the use of compacted utterances; while *articulation* might be associated with 'intelligibility'.

This remains for the moment an exploratory approach (but see Field 2011). A good illustration of why caution may be necessary is to be found in the term *fluency*, widely used in descriptors and rating scales. Evidence of surface dysfluency might reflect test taker hesitation at almost any stage: conceptualisation (forming ideas) – organisation (sequencing ideas) – formulation (retrieving relevant grammar/vocabulary/functional language) – encoding (forming a phonetic plan, embedding words in an easy-to-produce chunk of language) – articulation (forming the appropriate instructions to articulators). For authoritative discussions of the complex nature of 'fluency', see Segalowitz (2010) and Lickley (2015); for evidence of the effects of conceptualisation upon it, see Felker, Klockmann and De Jong (2019).

Well calibrated?

Item analysis of the kind just described can be extended to examine how the cognitive demands of a suite of tests vary across proficiency levels. Mention has been made of the fact that early-stage L2 learners have little spare attention to handle complex operations such as constructing or following a line of argument. On this basis, it will be evident that reading and listening items designed for lower proficiency levels (say up to CEFR B1+) should largely focus on factual reporting which engages the first three operations in Figure 1. At higher proficiency levels, items should, at least in part, engage the higher interpretive and integrative processes represented by other operations. This might take the form at B2 level of many more items targeting meaning construction (inference, speaker intentions, etc.); and at C levels of many more that require the reader/listener to report on the text as a whole.

A familiar factor determining difficulty in both receptive and productive skills is *discourse type.* Across all four skills, test designers and item writers tend to grade it as follows:

Lower levels (say A1–B1)	**Middle levels (say B2)**	**Higher levels (C1–C2)**
Narrative	Process-oriented	Complex exposition
Descriptive	Expository	Argument based
Instructional		Analytical
Informational		Persuasive

These choices are generally instinctive – based upon current syllabuses, perceived difficulty and vocabulary considerations. But they are strongly

endorsed if account is taken of the cognitive processes engaged. The discourse types in the first column are (speaking generally) simpler to process. Narrative and instructional texts in particular are usually sequential, so the connections between sentences or utterances are relatively transparent, even if no discourse markers are used. This makes them easier to follow for both readers and listeners, and easier to assemble under pressures of time by speakers. The relative simplicity of connections also assists the 'organisation' stage of writing.

Some of these basic types may still be in evidence at B2 level; but the increasing use of exposition there requires more complex connections and an awareness of the discourse markers involved. It also, as noted previously, reflects the fact that learners at this level become more capable of handling operations at the level of meaning. At C levels, the text types employed should require candidates to trace lines of argument across part or all of a text. Some caution may be necessary at the highest level (C2), where item writers are also sometimes exhorted to increase difficulty by using texts or productive tasks that are more 'abstract'. The term tends to be loosely defined and hard to apply. The resulting texts have sometimes been found to be unnaturally hard to follow, even for an L1 listener/reader (Field 2013:123–124).

Future directions

The goal of a cognitive validation exercise is to investigate what might be taking place in the mind of the test taker and comparing it to real-world behaviour. Given this, an increased use of verbal report might seem a sensible future step in order to supplement evidence derived from examining test material. However, obtaining reliable protocols is not easy. The reporting cannot take the form of a stream of consciousness account that interrupts the natural application of the skill under investigation – a point made often in relation to studies of writing. A possible alternative, *post hoc* reporting, is heavily reliant on memory.

Also daunting are the problems associated with trying to match a test taker's randomly reported behaviour against a model like those in Figures 1 and 2. As already noted, some operations are indispensable to the use of a skill (the example cited was the perceptual processes involved in reading) and therefore unhelpful if reported. Others are so closely integrated that the participant is unlikely to report on them (e.g. the speaking processes of turning an idea into words, storing the words in the mind and turning the words into instructions to the articulators). A further point is that cognitive processes are often automatic and therefore not available to report in the way that metacognitive ones are.[2] Participants are thus more likely to report

2 Hence the questionnaire results favouring metacognitive strategies that are often reported by commentators (e.g., Vandergrift 2003).

higher-level processes such as making use of context, drawing inferences and making logical connections than to report perceptual ones.

There is certainly a part to be played by verbal report; but it is perhaps most effective when used to shed light on how a given test shapes test taker behaviour. Findings of this kind can then be contrasted with the behaviour of real-world language users. Conclusions might include:

- how test conditions and formats affect performance
- correlations between high scores and certain aspects of performance
- how test takers deal with items targeting higher-level operations
- how candidates deal with gaps of knowledge or understanding by making use of compensatory strategies
- how candidates exploit loopholes of test design by employing test-wise strategies.

Once again, the receptive skills are easier to handle. The availability of a set of comprehension items lends itself well to a process of stimulated recall (*Why did you choose that answer? What do you think you heard/read?*). However, the usual provisos associated with any recall method must be applied. The lapse of time between locating an answer and reporting the process involved should be controlled to ensure reliability; this may entail breaking a listening or reading task into shorter sections that only cover the participant's responses to (say) three items. An alternative method is to rely on note-taking followed by an oral summary in L1 or L2. For examples of listening protocols obtained in this way, see Badger and Yan (2012) and Field (2012).

Investigating the productive skills in this fashion cannot be quite so transparent. For speaking, the most productive approaches entail recording the exchanges between examiner and test taker and replaying critical moments afterwards a) for the test taker to explain their rationale or b) for the examiner to comment in relation to the score awarded. The most productive approaches for writing entail halting the writing process at intervals for the writer to report what is in their mind. In terms of both skills, comments can be mapped against the operations shown in Figure 2.

Present and future investigations of cognitive validity are also fortunate in being able to draw upon important advances in technology (see the later chapter by Chan and Latimer). One thinks particularly of the availability of eye-tracking (Conklin, Pellicer-Sánchez and Carrol 2018, Godfroid 2019, Godfroid, Winke and Gass (Eds) 2013) and keystroke logging (Sullivan and Lindgren (Eds) 2006), which provide a detailed coded record of test taker performance.

These innovative approaches potentially offer exciting insights into specific cognitive areas. In relation to writing, keystroke logging can demonstrate how, under test conditions, the task of a writer is distributed between conceptualising, executing and revising (Lindgren and Sullivan 2006, Spelman

Miller 1996, 2000, Wengelin et al 2009). It can also be used to compare the behaviour entailed in computer- as against paper-based delivery of writing tasks (Barkaoui 2016), and the different patterns of writing behaviour seen in low-scoring as against more successful candidates.

In relation to tests of reading, eye-tracking can show how the attention of a candidate shifts from item to text and back again, thus providing insights into the impact of test method on performance. (This, of course, provides a picture of test performance but does not represent what occurs in real-world reading.) More generally, the technology can be used to compare the behaviour of L2 and L1 readers through evidence of a reader's interaction with the text in terms of fixations, saccades and regressions of the eye.

At this point, a note of caution should perhaps be sounded when considering issues of cognitive validity. Keystroke logging and eye-tracking record in impressive (sometimes excessive) detail what test takers actually do – but that is not the same as providing reliable evidence of the mental processes lying behind the activity (Holmqvist et al 2011:71–75). It is not easy to map from a visual record or from performance-related statistical data to the types of operation shown in Figures 1 and 2. Any process of interpretation needs to be rigorous and well supported. Unsurprisingly, it has become common practice to supplement the results of computational analysis with other sources of evidence, such as stimulated recalls supplied by participants.

Whatever its limitations, eye-tracking certainly offers an interesting means of investigating an issue dear to the heart of Cyril Weir – the use of different reading styles (expeditious vs careful, local vs global). For an exploratory study, see Bax and Weir (2012).

Final comments

This account has provided an overview of a strand of test validation that Cyril Weir was instrumental in bringing to the attention of specialists in language testing. In the process, he demonstrated his customary commitment to interdisciplinary exchanges of ideas; and he himself explored the writing and reading skills (Shaw and Weir 2007, Khalifa and Weir 2009, respectively) on the basis of such ideas.

As defined here, cognitive validation addresses the important issue of whether a language test can claim to be a predictor of the real-world performance of its candidates. It does so by matching certain test characteristics against empirically supported models of the four skills. In addition, taking due account of the role of automatic processing in skilled language use, it enables commentators to define levels of proficiency in a way that is informed by principle as well as (*pace* the CEFR) experience. In short, it meets the two criteria laid down by Weir (2005) as central to test validation. It operates in a *post hoc* manner in that it enables an assessment of the effectiveness

of existing tests that is not purely led by scores. It also operates *ab initio* by providing test designers with a framework for each skill and indicators of its possible distribution across levels of performance.

References

Anderson, J R (1983) *The Architecture of Cognition*, Cambridge: Harvard University Press.

Anderson, J R (2000) *Cognitive Psychology and its Implications*, New York: W H Freeman (Fifth edition).

Badger, R and Yan, X (2012) The use of tactics and strategies by Chinese students in the listening component of IELTS, in Taylor, L and Weir, C J (Eds) *IELTS Collected Papers 2: Research in Reading and Listening Assessment,* Studies in Language Testing volume 34, Cambridge: UCLES/Cambridge University Press, 454–486.

Barkaoui, K (2016) *Examining the cognitive processes engaged by Aptis Writing Task 4 on paper and on the computer*, London: British Council, available online: www.britishcouncil.org/sites/default/files/barkaoui.pdf

Bax, S (2013) The cognitive processing of candidates during reading tests: Evidence from eye-tracking, *Language Testing* 30 (4), 441–465.

Bax, S and Weir, C J (2012) Investigating learners' cognitive processes during a computer-based CAE Reading test, *Research Notes* 47, 3–14.

Baxter, G P and Glaser, R (1998) Investigating the cognitive complexity of science assessments, *Educational Measurement: Issues and Practice* 17 (3), 37–45.

Bereiter, C and Scardamalia, M (1987) *The Psychology of Written Composition,* New Jersey: Lawrence Erlbaum Associates, Inc.

Brunfaut, T and McCray (2015) *Looking into test-takers' cognitive processes while completing reading tasks: a mixed-method eye-tracking and stimulated recall study*, London: British Council, available online: www.britishcouncil.org/sites/default/files/brunfaut-and-mccray-report_final.pdf

Cain, K and Oakhill, J (Eds) (2008) *Children's Comprehension Problems in Oral and Written Language: A Cognitive Perspective*, New York: Guilford Press.

Chan, S (2018) *Defining Integrated Reading-into-Writing Constructs: Evidence at the B2–C1 Interface,* English Profile Studies volume 8, Cambridge: UCLES/Cambridge University Press.

Conklin, K, Pellicer-Sánchez, A and Carrol, G (2018) *Eye-Tracking: A Guide for Applied Linguistics Research,* Cambridge: Cambridge University Press.

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

Ellis, N C (1994) Implicit and explicit language learning: an overview, in Ellis, N C (Ed) *Implicit and Explicit Learning of Languages*, Amsterdam: Academic Press, 1–31.

Ellis, R, Loewen, S, Elder, C, Erlam, R, Philp, J and Reinders, H (2009) *Implicit and Explicit Knowledge in Second Language Learning*, *Testing and Teaching*, Bristol: Multilingual Matters.

Ericsson, K A and Smith, J (Eds) (1991) *Toward a General Theory of Expertise: Prospects and Limits,* Cambridge: Cambridge University Press.

Felker, E R, Klockmann, H E and De Jong, N (2019) How conceptualizing influences fluency in first and second language speech production, *Applied Psycholinguistics* 40 (1), 111–136.

Feltovich, P J, Prietula, M J and Ericsson, K A (2006) Studies of expertise from psychological perspectives, in Ericsson, K A, Charness, N, Feltovich, P J and Hoffman, R R (Eds) *The Cambridge Handbook of Expertise and Expert Performance*, Cambridge: Cambridge University Press, 41–67.

Field, J (1999) Key concept: bottom-up and top-down, *ELT Journal* 5 (4), 338–339.

Field, J (2004) *Psycholinguistics: The Key Concepts,* London: Routledge.

Field, J (2011) Cognitive validity, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Second Language Speaking,* Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press, 65–111.

Field, J (2012) The cognitive validity of the lecture-based question in the IELTS listening paper, in Taylor, L and Weir, C J (Eds) *IELTS Collected Papers 2: Research in Reading and Listening Assessment,* Studies in Language Testing volume 34, Cambridge: UCLES/Cambridge University Press, 391–453.

Field, J (2013) Cognitive validity, in Geranpayeh, A and Taylor, L (Eds) *Examining Listening: Research and Practice in Assessing Second Language Listening,* Studies in Language Testing volume 35, Cambridge: UCLES/ Cambridge University Press, 77–151.

Field, J (2015) Psycholinguistics, in Braber, N, Cummings, L and Morrish, L (Eds) *Exploring Language and Linguistics*, Cambridge: Cambridge University Press, 324–352.

Field, J (2018a) The cognitive validity of tests of listening and speaking designed for young learners, in Papp, S and Rixon, S, *Examining Young Learners: Research and Practice in Assessing the English of School-age Learners*, Studies in Language Testing volume 47, Cambridge: UCLES/Cambridge University Press, 128–200.

Field, J (2018b) The cognitive validity of tests of reading and writing designed for young learners, in Papp, S and Rixon, S, *Examining Young Learners: Research and Practice in Assessing the English of School-age Learners*, Studies in Language Testing volume 47, Cambridge: UCLES/Cambridge University Press, 201–269.

Field, J (2019) *Rethinking the Second Language Listening Test*, Sheffield: Equinox.

Gagnon, R, Charlin, B, Roy, L, St-Martin, M, Sauvé, E, Boshuizen, H P and van der Vleuten, C (2006) The cognitive validity of the script concordance test: A processing time study, *Teaching and Learning in Medicine* 18 (1), 22–27.

Gilmore, A (2015) The influence of discourse studies on language descriptions and task design, *Language Teaching* 48 (4), 506–530.

Glaser, R (1991) Expertise and assessment, in Wittrock, M C and Baker, E L (Eds) *Testing and Cognition*, Englewood Cliffs: Prentice Hall, 17–30.

Godfroid, A (2019) Investigating instructed second language acquisition using L2 learners' eye-tracking data, in Leow, R P (Ed) *The Routledge Handbook of Second Language Research in Classroom Learning*, Oxford: Routledge, 44–57.

Godfroid, A, Winke, P M and Gass, S (Eds) (2013) Thematic issue on eye tracking in second language acquisition research, *Studies in Second Language Acquisition* 35.

Goldman, S R (2004) Cognitive aspects of constructing meaning through and across multiple texts, in Shuart-Ferris, N and Bloome, D (Eds) *Uses of Intertextuality in Classroom and Educational Research*, Greenwich: Information Age Publishing, 317–351.

Grellet, F (1981) *Developing Reading Skills,* Cambridge: Cambridge University Press.

Hartman, D K (2005) Eight readers reading: The intertextual links of proficient readers reading multiple passages, *Reading Research Quarterly* 30 (3), 520–561.

Hayes, J R and Flower, L S (1986) Writing research and the writer, *American Psychologist* 41 (10), 1,106–1,113.

Holmqvist, K, Nyström, M, Andersson, R, Dewhurst, R, Jarodzka, H and van de Weijer, J (2011) *Eye Tracking*, Oxford: Oxford University Press.

Holzknecht, F and Eberharter, K (2017) *Looking into listening: Using eye-tracking to establish the cognitive validity of the Aptis Listening Test*, London: British Council, available online: www.britishcouncil.org/sites/default/files/looking_into_listening.pdf

Kellogg, R T (1994) *The Psychology of Writing*, Oxford: Oxford University Press.

Kellogg, R T (1996) A model of working memory in writing, in Levy, M C and Ransdell, S (Eds) *The Science of Writing*, New Jersey: Lawrence Erlbaum Associates, Inc., 57–71.

Khalifa, H and Weir, C J (2009) *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing volume 29, Cambridge: UCLES/Cambridge University Press.

Koskey, K L, Karabenick, S A, Woolley, M E, Bonney, C R and Dever, B V (2010) Cognitive validity of students' self-reports of classroom mastery goal structure: What students are thinking and why it matters, *Contemporary Educational Psychology* 35 (4), 254–263.

Lacroix, N (1999) Macrostructure construction and organisation in the processing of multiple text passages, *Instructional Science* 27, 221–233.

Larichev, O I (1992) Cognitive validity in design of decision-aiding techniques, *Journal of Multi-Criteria Decision Analysis* 1, 127–138.

Levelt, W J M (1989) *Speaking: From Interaction to Articulation*, Cambridge: MIT Press.

Levelt, W J M (1999) Producing spoken language: a blueprint of the speaker, in Brown, C and Hagoort, P (Eds) *The Neurocognition of Language*, Oxford: Oxford University Press, 83–121.

Lickley, R (2015) Fluency and dysfluency, in Redford, M A (Ed) *The Handbook of Speech Production,* Chichester: Wiley, 445–469.

Lindgren, E and Sullivan, K P H (2006) Writing and the analysis of revision, in Sullivan, K P H and Lindgren, E (Eds) *Computer Keystroke Logging and Writing: Methods and Applications,* Amsterdam: Elsevier, 31–44.

Linn, R L, Baker, E L and Dunbar, S B (1991) Complex performance-based assessment: Expectations and validation criteria, *Educational Researcher* 20 (8), 5–21.

MacArthur, C A, Graham, S and Fitzgerald, J (Eds) (2006) *Handbook of Writing Research*, New York: Guilford Press.

McCray, G and Brunfaut, T (2016) Investigating the construct measured by banked gap-fill items: Evidence from eye tracking, *Language Testing* 35 (1), 1–23.

Messick, S (1989) Validity, in Linn, R L (Ed) *Educational Measurement*, New York: American Council on Education/Macmillan (Third edition), 13–103.

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

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

Oakhill, J and Garnham, M (1988) *Becoming a Skilled Reader,* London: Blackwell.

Pashler, H and Johnston, J C (1998) Attentional limitations in dual task performance, in Pashler, H (Ed) *Attention,* Hove: Psychology Press, 155–189.
Perfetti, C A, Rouet, J and Britt, M A (1999) Toward a theory of document representation, in van Oostendorp, H and Goldman, S R (Eds) *The Construction of Mental Representations during Reading,* London: Lawrence Erlbaum Associates, Inc., 99–122.
Rayner, K, Pollatsek, A, Ashby, J and Clifton, C (2012) *Psychology of Reading*, New York: Psychology Press (Second edition).
Rebuschat, P (Ed) (2015) *Implicit and Explicit Learning of Languages*, Amsterdam: John Benjamins.
Richards, J (1983) Listening comprehension: approach, design, procedure, *TESOL Quarterly* 17, 219–239.
Ruiz-Primo, M A and Shavelson, R J (1996) Rhetoric and reality in science performance assessments: an update, *Journal of Research in Science Teaching* 33 (10), 1,045–1,063.
Ruiz-Primo, M A, Schultz, S E, Li, M and Shavelson, R J (2001) On the validity of cognitive interpretations of scores from alternative concept-mapping techniques, *Educational Assessment* 7 (2), 99–141.
Scardamalia, M and Bereiter, C (1987) Knowledge telling and knowledge transforming in written composition, in Rosenberg, S (Ed) *Advances in Applied Psycholinguistics: Volume 2,* Cambridge: Cambridge University Press, 142–175.
Segalowitz, N (2010) *Cognitive Bases of Second Language Fluency,* New York: Routledge.
Segalowitz, N (2013) Automaticity, in Robinson, P (Ed) *The Routledge Encyclopaedia of Second Language Acquisition,* Abingdon: Routledge, 53–57.
Shaw, S D and Weir, C J (2007) *Examining Writing: Research and Practice in Assessing Second Language Writing*, Studies in Language Testing volume 26, Cambridge: UCLES/Cambridge University Press.
Snow, R E, Corno, L and Jackson, D (1996) Individual differences in affective and conative functions, in Berliner, D C and Calfee, R C (Eds) *Handbook of Educational Psychology,* New York: Macmillan, 243–310.
Snowling, M and Hulme, C (Eds) (2005) *The Science of Reading: A Handbook*, Oxford: Blackwell.
Spelman Miller, K (1996) The pausological study of written language production, in Sullivan, K P H and Lindgren, E (Eds) *Computer Keystroke Logging and Writing: Methods and Applications,* Amsterdam: Elsevier, 11–30.
Spelman Miller, K (2000) Academic writers on-line: investigating pausing in the production of text, *Language Teaching Research* 4 (2), 123–148.
Strømsø, H I and Bråten, I (2002) Norwegian law students' use of multiple sources while reading expository texts, *Reading Research Quarterly* 37 (2), 208–237.
Sullivan, K P H and Lindgren, E (Eds) (2006) *Computer Keystroke Logging and Writing*, Amsterdam: Elsevier.
Thelk, A D and Hoole, E R (2006) What are you thinking? Postsecondary student think-alouds of scientific and quantitative reasoning items, *Journal of General Education* 55 (1), 17–39.
Torrance, M and Jeffery, G (1999) Writing processes and cognitive demands, in Torrance, M and Jeffery, G (Eds) *The Cognitive Demands of Writing: Processing Capacity and Working Memory Effects in Text Production*, Amsterdam: Amsterdam University Press, 1–11.

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

Vandergrift, L (2003) Orchestrating strategy use: Toward a model of the skilled second language listener, *Language Learning* 53, 463–496.

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

Weir, C J and Chan, S (2019) *Research and Practice in Assessing Academic Reading: The Case of IELTS*, Studies in Language Testing volume 51, Cambridge: UCLES/Cambridge University Press.

Wengelin, A, Torrance, M, Holmqvist, K, Simpson, S, Galbraith, D, Johansson, V and Johansson, R (2009) Combined eye-tracking and keystroke-logging methods for studying cognitive processes in text production, *Behavior Research Methods* 41 (2), 337–351.

Westby, C (2004) A language perspective on executive functioning, metacognition and self-regulation in reading, in Stone, C, Silliman, E R, Ehren, B J and Apel, K (Eds) *Handbook of Language and Literacy,* New York: Guilford Press, 398–427.

Yuill, N and Oakhill, J (1991) *Children's Problems in Text Comprehension: An Experimental Investigation,* Cambridge: Cambridge University Press.

4 Context validity in language assessment: test operations and conditions for construct operationalisation

Yan Jin
Shanghai Jiao Tong University

This chapter is a tribute to Professor Cyril J Weir, whose long involvement with language tests and language testers in mainland China has contributed significantly to the growth and development of professionalism and expertise in the field of language assessment in the country. To be specific, the chapter aims to demonstrate the value of Weir's (2005) socio-cognitive framework for operationalising the constructs of two locally developed language tests. To this end, the chapter is devoted to:

- a review of the role of context validity in test development and validation
- an analysis of two cases in which test constructs were maximally operationalised
- a reflection on lessons learned from the two case studies, and finally
- a discussion of the challenges in establishing context validity.

Introduction

In the early 1990s, Dr Cyril J Weir, then a senior lecturer at the Centre for Applied Language Studies (CALS), University of Reading, was appointed by the British Council as a consultant to the National College English Testing Committee in China to work on the validation of the College English Test (CET) (Jin 2010, Yang and Weir 1998, Zheng and Cheng 2008). One of the tasks the project team was assigned was to specify 'operations' and 'conditions' for each item or task of the test, an activity Davidson and Lynch (2002:41) would call 'reverse engineering' of specification-driven testing. Specifying test operations and performance conditions was not standard practice among Chinese practitioners about three decades ago. Expert judgement was often used as evidence for *post hoc* evaluation of the correspondence between an item/task and the knowledge, skills, and abilities that the

item/task was intended to measure. The evidence was collected to validate what was called content validity, one of 'the earlier trio of validities' (Chapelle 1999:256), which is concerned with the 'content relevance, representativeness, and technical quality' of the test material (Messick 1995:6).

This concern with content validity represents a communicative paradigm of language testing advocated in Weir (1990, 1993). As noted in Weir (1990:7), communicative language ability is tested by 'evaluating samples of performance, in certain specific contexts of use, created under particular test constraints'. Such a communicative approach was vividly exemplified in the development of the Test in English for Educational Purposes (TEEP). Weir (1990) provided useful guidelines for establishing content validity: examining systematically the behaviour domain; describing fully the domain under consideration before test development; and investigating the relevance of the individual's test responses to the behaviour area under consideration. In the companion volume of Weir's 1990 work, Weir (1993:28) further explained: '[I]f the test tasks reflect real-life tasks in terms of important identified conditions and operations it is easier to state what a student can do through the medium of English'.

The importance of assessment context for eliciting high-quality communicative performance was explicitly mentioned and clearly demonstrated in both volumes but not foregrounded and overtly articulated until the early 2000s when the socio-cognitive validity theory was proposed (Weir 2005). In the theoretical model for validating language assessments, context validity is conceptualised as an essential component that is concerned with 'the extent to which the choice of tasks in a test is representative of the larger universe of tasks of which the test is assumed to be a sample' (Weir 2005:19). Contextual facets, from a socio-cognitive perspective, fulfil a bridging role between the construct to be assessed and the communicative competence required in real-world situations.

In this chapter, the concept of context validity is revisited and the central role of contextually appropriate operations and conditions in construct operationalisation is exemplified through a detailed analysis of two cases: expeditious reading tasks in the Advanced English Reading Test (AERT) and the peer-to-peer discussion in the CET-Spoken English Test (CET-SET). The two cases were chosen for their salient contextual features in assessing the receptive skill of reading and the productive skill of speaking. The chapter ends with reflections on the lessons learned from the two case studies and the challenges in extrapolating from test performances to performances in target language use situations. It also acknowledges Cyril Weir's considerable legacy as far as the development of professional expertise and experience in the field of language assessment is concerned in mainland China.

Context validity revisited

In this section, the concept of context validity is reviewed to highlight the place of contextual facets in operationalising assessment constructs. Particular references are made to the application of the socio-cognitive framework in Cambridge English language examinations (Weir, Vidaković and Galaczi 2013), Cambridge English assessments of reading (Khalifa and Weir 2009), and Cambridge English assessments of speaking (Taylor (Ed) 2011).

The context of language use

The value of context for appropriate use of language began to be recognised in the early 1970s when communicative competence was defined for language teaching and testing. Hymes (1972) proposed a communicative competence model which comprised two dimensions: a linguistic dimension, or the knowledge underlying linguistic performance, and a sociolinguistic dimension, i.e. the ability to use the language in specified contexts. Canale and Swain (1980) and Canale (1983) further developed the theory of communicative competence with a four-dimension model, which made a clear distinction between sociolinguistic competence and discoursal competence. The communicative competence theory of direct relevance to language testing was the Communicative Language Ability (CLA) model put forward in Bachman (1990) and refined in Bachman and Palmer (1996). The model expanded the scope of the construct of communicative competence and called for attention to test methods and task characteristics. Weir (1990, 1993) and other language testers in the UK were also committed to the communicative approach to language testing and sought to 'identify those abilities (operations/activities) and performance conditions that seemed to be important components of language use in particular contexts' (Weir 2013:79).

The socio-cognitive validity theory was a result of the attempt to provide useful guidance on assessing communicative language competence. Weir (2013:3) commented on the socio-cognitive approach adopted by Cambridge ESOL (now Cambridge Assessment English) to construct validation: 'this approach is effectively an interactionalist position, which sees the construct as residing in the interaction between an underlying cognitive ability, a context of use and a process of scoring'. By adopting such an approach, the abilities are demonstrated by the cognitive processing of the candidate, and the use of language in performing tasks is viewed as 'a social rather than a purely linguistic phenomenon' (Weir 2013:3). Contextual facets, therefore, constitute one of the core components of a test's validity. Taylor and Galaczi (2011:172) used the term 'the core of the socio-cognitive framework' to refer to the three key dimensions of the socio-cognitive framework: cognitive validity, context validity and scoring validity – 'at the heart of any language testing activity . . .

we can conceive of a triangular relationship between three essential components: the test taker's cognitive abilities; the task and context; and the scoring process'.

The context of language use, as discussed above, plays a central role in assessing communicative competence. According to Weir (2013:3), two fundamental questions should be addressed in order to establish the context validity of a language test:

1. Are the characteristics of the test task an adequate and comprehensive representation of those that would be normally encountered in the real-life context?
2. Are they appropriately calibrated to the level of proficiency of the learner being evaluated?

The two questions are discussed in some detail below with specific focuses on reading and speaking assessments in the hope of guiding the analysis of the two cases in the following section.

Contextual features of reading assessments

Weir's (2005:44) framework for validating reading tests takes into consideration three sets of contextual parameters that are likely to have an impact on performances on reading tests: task setting, linguistic demands and administrative setting (see also Khalifa and Weir 2009:Chapter 4).

Task-setting parameters are most directly relevant to the design of assessment tasks that best simulate communicative activities in real-life contexts. When the construct of an assessment task is defined, the most important decision for assessment design is task format or response method. There are basically two types of formats for assessing reading: selected responses (e.g. multiple-choice questions, true/false items, matching) and constructed responses (e.g. short-answer questions, gap filling, reading into writing). By adopting different task formats, test takers will be engaged in a variety of cognitive processes that tap into different aspects of the construct of reading. Other considerations of the context of reading tasks include weighting of tasks, test takers' knowledge of scoring criteria, order of items, channel of presentation (verbal or non-verbal), text length, and time constraints.

In addition to task setting, linguistic demands should also be specified in order to elicit performances on reading tasks at an appropriate level of difficulty. The linguistic knowledge for reading tasks can be described in terms of overall text purpose (text types), writer–reader relationship, discourse mode (genre, rhetorical task, pattern of exposition), functional resources, grammatical resources, lexical resources, nature of information (abstract or concrete), and content knowledge (including topical knowledge). Administrative setting is concerned with the circumstances under which a task is performed

in an assessment. For reading tasks, administrative parameters include mainly physical conditions (e.g. venues, seating arrangements), uniformity of administration (e.g. time control, supervision), and security (confidentiality of test materials).

Examples were provided in Khalifa and Weir (2009) to show how these contextual facets of reading tasks are manipulated to better represent the construct of reading and target at different levels of reading proficiency in Cambridge English (or General English) examinations. Take text length as an example. The examinations in this suite are KET, PET, FCE, CAE and CPE, aligned to the CEFR Levels A2, B1, B2, C1, and C2 respectively.[1] The length of the text used in each test increases as the examination level increases (Khalifa and Weir 2009:101): four texts in KET, each maximally 250 words, totalling 740–800 words; five texts in PET, each maximally 550 words, totalling 1,450–1,600 words; three texts in FCE, each maximally 700 words, totalling 2,000 words; six texts in CAE, each maximally 1,100 words, totalling 3,000 words; and nine texts in CPE, each maximally 1,100 words, totalling 3,000 words. A similar gradation in textual complexity can be seen in the average difficulty estimates of the texts (Flesch reading ease score/Flesch-Kincaid grade level) in these examinations (Khalifa and Weir 2009:122): KET 78.3/5.5; PET 64.7/7.9; FCE 66.5/8.4; CAE 58.4/9.6; and CPE 57.7/9.9.

Contextual features of speaking assessments

Similar to the contextual variables for assessing reading, Weir's (2005:46) socio-cognitive framework considers three sets of parameters for validating speaking tests: task setting, linguistic demands, and administrative setting (see also Taylor and Galaczi 2011).

The most important task-based parameter for speaking tests is task format, which has proved to affect test takers' speaking performances. Speaking assessments typically employ monologic or dialogic tasks. When monologic tasks are performed, test takers are engaged in such activities as reading aloud, answering questions, or making a presentation/speech. Dialogic tasks represent the interactive nature of oral communication, in which test takers are engaged in pair or group discussions with the examiner or peer candidate(s). While the benefits of pair or group work in assessment contexts are recognised, the implementation of interactive tasks is challenging due to the co-constructed nature of interactive discourses. Much attention has been paid in the literature to issues of the joint construction of

1 The Cambridge English exams have now been rebranded as A2 Key, B1 Preliminary, B2 First, C1 Advanced, and C2 Proficiency (www.cambridgeenglish.org/exams-and-tests/qualifications/general/).

performance between test takers (e.g. McNamara 1997, see also the special issue of *Language Testing* edited by Taylor and Wigglesworth 2009).

Other task-based considerations of speaking assessments include knowledge of criteria, weighting, order of items/tasks, and time constraints. The condition of 'knowledge of criteria' is less relevant to selected-response tasks often used in reading assessments. Speaking assessments are performance-based and rating scales are *de facto* constructs. Test takers, therefore, should be provided in advance with as much information as possible about rating criteria. Test takers' knowledge of the criteria will affect their performances on tasks. It is also interesting to note that in speaking assessments, time constraints refer to the time for planning as well as task completion. Research has suggested that planning time interacts with speakers' level of proficiency and task requirements, so the effect of planning differs for speakers at different proficiency levels or engaged in tasks of varying cognitive demands.

Apart from task-based features, the demands of speaking tasks also relate to their linguistic input and output. The parameters include channel of communication (aural, written, visual and graphical), discourse mode (narration, description, exposition, and argument/persuasion), length of input, nature of information (abstract or concrete), topic familiarity/content knowledge required, and lexical, structural and functional resources. Among other contextual parameters of oral performances, interlocutor variables are of particular importance to the context validity of oral interaction tasks. Weir (2005) identified the following interlocutor variables for designing and analysing speaking tasks: speech rate, variety of accent, acquaintanceship, number, and gender. O'Sullivan (2002) added to the list several parameters contributing to interlocutor effects: age, cultural background, proficiency level, personality, and conversation style.

Examples were provided in Galaczi and ffrench (2011) to show how these contextual facets of speaking tests have been manipulated to best represent the construct of speaking in the Cambridge English examinations. Nature of information, for example, is considered as 'one of the most salient features in determining the difficulty of tasks across the five Main Suite levels' (Galaczi and ffrench 2011:145–149). There is a progression from mostly personal information at the lower levels of A2 to B2 (KET, PET, FCE) to mostly non-personal information at the higher levels of C1 and C2 (CAE, CPE). The degree of concreteness or abstractness of the information given in the input or elicited by the task can also be manipulated to control task difficulty. Tasks of lower proficiency levels (KET, PET) elicit mostly concrete or factual information, whereas for test takers at the highest level of proficiency (CPE), responses comprise mostly abstract information. In between, both concrete and abstract information can be found in FCE and CAE.

Context validity in action: two case studies

In this section, two case studies, the Advanced English Reading Test (AERT) and the CET Spoken English Test (CET-SET), are analysed to demonstrate the importance of specifying test operations and performance conditions for establishing the context validity of language assessments. Lessons learned from the two case studies for future development of reading and speaking assessments are summarised in the next section.

Case Study 1: assessing expeditious reading in the AERT

The development and validation of the AERT was a Sino-British co-operative project initiated by the National College English Testing Committee and supported by the British Council in the mid-1990s. Cyril Weir was the project consultant and co-supervisor of the author, who was working on the project for her doctoral dissertation. The construct of the test was defined based on a review of literature on the componentiality of reading ability, an analysis of Chinese readers' difficulties in reading in English as a foreign language, an analysis of Chinese undergraduates' English for Academic Purposes (EAP) reading needs, and an analysis of texts and tasks in EAP reading textbooks and tests (Jin 2002, Weir, Yang and Jin 2000).

Operations: expeditious reading

Weir (2005:18) stressed the importance of having an expeditious reading component in reading assessments so as to ensure construct representativeness: 'If we only test careful reading and not expeditious reading (see Urquhart and Weir 1998), are we measuring all of reading ability?' The AERT considers 'expeditious reading' as an essential component of the test's constructs and defines it as 'the process of reading the text *selectively* without necessarily following the linearity of the text' (Jin 2002:207). Instead of reading speed, the emphasis is laid on the efficiency in achieving the purposes for reading. While reading expeditiously, readers are expected to consciously use various strategies to sample the text to be read. The three strategies identified for Chinese EAP readers are skimming, search reading, and scanning.

Skimming is reading at the global level for the discourse topic of a long text. The construct of skimming can be operationalised by using the following strategies:

- reading the title and subtitles quickly
- reading the abstract carefully
- reading the introductory and the concluding paragraphs carefully
- reading the first and the last sentences of each paragraph carefully
- glancing at words and phrases.

Search reading is also a type of reading at the macrostructural level. The purpose of search reading is to locate information in the macrostructure of the text on predetermined topics. The construct of search reading can be operationalised through the use of the following strategies:

- keeping alert for words in the same or related semantic field
- using formal knowledge for locating information
- using titles and subtitles
- reading abstracts where appropriate
- glancing at words or phrases.

Scanning is fast reading at the local level for the purpose of locating specific symbols, numbers, dates and so on in the microstructure of the text. The construct can be operationalised through the use of the strategy of matching of specific words, phrases, figures, numbers, dates and names.

Conditions: expeditious reading

For an assessment of reading, the operationalisation of the construct depends to a large extent on tasks, texts, and the implementation of the assessment. When the AERT specifications were developed, efforts were made to fully specify task-based conditions, or in Weir's (2005) term, task-setting parameters (e.g. response method, speed of processing, amount of help provided, number and ordering of tasks, rubrics, weighting). Table 1 summarises the task-based conditions of the expeditious reading tasks in the AERT.

Table 1 Task-based conditions: expeditious reading in the AERT

	Skimming	Search reading	Scanning
Response method	Summarising the text in one sentence	Completing flow-chart, table and sentences (within eight words)	Completing flow-chart, table and sentences (single word or phrase)
Number and ordering of items	Three items, following the order of the three texts	12 items, following the sequence of the information in the text	15 items, following the sequence of the information in the text
Weighting	Equal weighting	Equal weighting	Equal weighting
Speed of processing	100–150 wpm*	100–150 wpm	100–150 wpm
Time constraints	15 minutes	21 minutes	18 minutes

**words per minute*

Text-based parameters contribute to the linguistic demands on the reader and are regarded as the main factors for determining the difficulty level of a reading task. Texts for skimming and search reading, for example, should

have an overt structure with a clear line of argument. As texts involving problem-solving, causation and comparison have clearly organised structures, they were considered the most suitable for assessing the two constructs. The skimming and searching sections of the AERT therefore share the same three texts. The scanning section uses another three texts, which have an explicit structure and are of a descriptive nature. Table 2 summarises the text-based conditions of the expeditious reading tasks in the AERT.

Table 2 Text-based conditions: expeditious reading in the AERT

	Skimming and search reading	**Scanning**
Text length	Three texts, each c.1,000 words	Three texts, each c.1,000 words
Source of texts	Journal articles, chapters of textbooks	Journal articles, chapters of textbooks
Illocutionary features	Inform, describe, explain	Describe
Level of language difficulty	Low to medium	Low to medium
Level of topic familiarity	Medium to high	Medium to high
Level of subject specificity	Low to medium	Low to medium
Rhetorical organisation	Causation, comparison, problem-solving	Collection of descriptions
Content knowledge	One text on each topic area: arts and humanities, science and technology, medical and life sciences	One text on each topic area: arts and humanities, science and technology, medical and life sciences

Research and validation

Following the specification of task operations and performance conditions, the next step is to 'operationalise as many of these parameters as faithfully as possible in the test task(s)' (Khalifa and Weir 2009:81). In the AERT project, research was undertaken to ensure that the operations and conditions were fulfilled in a principled and systematic way (Jin 2002).

To check the suitability of the texts selected for expeditious reading tasks, EAP teachers were surveyed for their perceptions of topic familiarity, subject specificity and language difficulty of each text (see Table 2 for the specifications of these conditions). Decisions on the texts were made based on the survey results. When items were developed, 'mind-mapping' was performed for selecting 'question areas', that is, the content of a text that should be extracted in line with the established purpose for reading. Targeted test takers mapped each text by reading expeditiously for an intended purpose (i.e. skimming, search reading, scanning). They highlighted the most salient parts in the text or recalled the content that

left a deep impression on them. Consensus on question areas was reached through group discussion. In addition, the procedure helped determine the time required for each task.

At the stage of assessment delivery, time was strictly controlled for each section of the test using coloured text and question booklets. Factor analysis of the item-level data, however, did not reveal a clear factorial structure of the constructs. The data of test takers' non-responses (blank answers) indicated that more time was spent on the first two texts than on the third text in each section. As strict control of response time at the text level was simply not possible in the paper-based mode of test delivery, the constructs of expeditious reading were found to be somewhat contaminated. Qualitative data of expert judgement, test taker introspection and retrospection were also collected. Typical performances in line with the purposes of skimming, search reading and scanning, as well as performances not expected by the test designer, were explicated for a better understanding of test takers' use of expeditious reading strategies for text processing and task completion.

Case Study 2: the peer-to-peer discussion task in the CET-SET

In the late 1980s when the CET was started, speaking was not an essential requirement for college students and was therefore not included in the test. In the late 1990s, with China's further opening up, the CET Spoken English Test (CET-SET) was developed to meet the social needs for university graduates with a high level of proficiency in English speaking. The test adopted the face-to-face format with three candidates and two examiners forming a group to complete a number of tasks including a peer-to-peer discussion task. The test gained popularity among university students. Until 2012, a total of 58 CET-SET test centres had been established in 35 major cities and over 1,000 CET-SET examiners had been trained and authorised across the country. But the scale of the face-to-face test was severely limited by the human resources required for test implementation as well as the need for training a larger number of qualified examiners. In 2013, the computer-based CET-SET replaced the face-to-face format. The computer-based oral test adopted a paired format: two candidates form a pair to complete a number of tasks including an online peer-to-peer discussion task.

Operations: the discussion task

In the late 1980s, while admitting the value of the communicative competence theory for language teaching and learning, scholars began to recognise the inadequacy of the theory to address interactional competence, that is, the knowledge, skills and strategies language users employ to bring about successful interaction. He and Young (Eds) (1998:5) noted that 'abilities, actions, and activities do not belong to the individual but are jointly constructed by all

participants'. Young (2008:101) defined interactional competence as 'a relationship between the participants' employment of linguistic and interactional resources and the contexts in which they are employed'. Weir (2005:72) also called for attention to the reciprocal nature of oral interaction: 'if we wish to test spoken interaction, a valid test must include reciprocity conditions'. He further noted that oral interaction 'contrasts with the traditional interview format in which the interviewer asks the questions and the interviewee answers' (2005:72).

When the CET-SET was developed, interactional competence was viewed as an essential component of the speaking constructs; hence, there was a strong rationale for having a discussion task in the test. The design underscored the need to assess reciprocal oral interaction by engaging test takers in a pair or group discussion. The CET-SET specifications (National College English Testing Committee 1999, 2016) describe the following operations for the oral interaction (discussion) task:

- exchanging ideas or views and expressing feelings or emotions
- engaging in debate or argument, giving explanations, and making comparisons
- using appropriate oral communication strategies to facilitate oral interaction.

In addition to task operations, the CET-SET specifications stipulate the following categories of 'interactional language functions' (ILFs) for the discussion task (see also He and Dai 2006:378–379):

1. Express (dis)agreement with what another speaker has said.
2. Ask for opinions or information.
3. Challenge opinions or assertions made by another speaker by giving countering reasons or evidence.
4. Support opinions or assertions made by another speaker by providing more reasons or evidence.
5. Modify arguments or opinions in response to another speaker.
6. Persuade another speaker to accept one's view.
7. Express ideas building on what another speaker has said.
8. Negotiate meaning: ask for clarification; give clarification; ask for confirmation; check for comprehension.

The constructs of interactional competence are also operationalised in the test via a rating scale consisting of three sets of criteria: accuracy and range, size and discourse management, and flexibility and appropriacy. The criteria of 'flexibility and appropriacy' are especially useful for assessing test takers' pragmatic competence: *flexibility* in dealing with various communicative situations and topics, and *appropriacy* in their use of linguistic resources

according to communicative contexts. These competencies are most likely to be elicited in an interactive task in which communication strategies are essential for completing tasks in a flexible manner and in appropriate language.

Conditions: the discussion task

The face-to-face discussion was set as a group task for three (or in rare cases four) test takers. The decision on a group format was made on the grounds of test efficiency given the scale of the CET-SET test. The test format was also used to balance out possible interlocutor effects (O'Sullivan 2002), which may potentially impact co-constructed discourses. In the group format, there could be more variability in interlocutors' level of oral proficiency and other background variables such as personality or conversation style.

In the transition from the face-to-face mode to the computer-based mode, the group format was replaced by a paired format, primarily due to the concern with the difficulty in distinguishing test takers' voices in the process of rating. The computer-based CET-SET is scored by raters who listen to test takers' recordings after the test. It is not easy for raters to distinguish and remember each test taker's voice in the discussion task. So the system is programmed to pair a boy student with a girl student until there are no more pairs of a different gender. In the paired format, it is less likely to balance out possible interlocutor effects. A possible way to mitigate interlocutor effects, especially the influence of the interlocutor's level of proficiency, is to choose topics that are familiar to test takers and relevant to their communicative needs. By so doing, test takers at different proficiency levels may all have something to say in the discussion.

To maximally operationalise the constructs of interactional competence, contextual parameters have been delineated for the CET-SET pair or group discussion task with a focus on interlocutor variables (see Table 3).

Table 3 Interlocutor-related features: the discussion task of the CET-SET

	Pair or group discussion in the CET-SET
Test taker background	No marked difference in linguistic, cultural and educational background
Speech rate and accent	May vary to some extent but within an acceptable range; mainly American or British accents, possibly with some local accents
Acquaintanceship	Get to know each other through self-introduction and previous tasks
Gender	Opposite gender preferred in computer-based pair discussion
Proficiency	Possible differences among group members
Personality	Possible differences among group members
Conversation style	Possible differences among group members

The CET-SET test takers, on the whole, form a homogenous group: they share similar linguistic, cultural and educational backgrounds. They are university students with Chinese as their first language. They have learned English for an average of 10–12 years. During the first two years in the university, they are required to take the College English course, which is a compulsory requirement for all university students. The speech rate and the accent of the CET-SET test takers may vary to some extent depending on their English learning experiences (British or American English) and the variety of their spoken Chinese, but the variability is within an acceptable range. Test takers in the pair or group do not know each other before taking the test, but they get to know each other through self-introduction, and they also have chances to become more familiar with each other through their performances on monologic tasks (e.g. question and answer, individual presentation). By the time they start the discussion, the group members have already had some idea of each other's oral proficiency levels, hobbies, views on some social issues, and so on. In the computer-based CET-SET, test takers can 'see' each other via the photos on the screen. Video cameras on the computer are not used in the test primarily due to the concern about bandwidth.

Contextual variables of the CET-SET have also been manipulated to control the difficulty level of the tasks (see Table 4). To contextualise its discussion tasks, the CET-SET sets clear communicative goals for the discussion task so that test takers have relevant roles to play and an authentic purpose of communication. In the CET-SET Band 4, the lower level of the test, the communicative goal is to reach an agreement through discussion on a concrete topic, such as a travel plan, the arrangement for an event, and so on. The prompts are mostly pictures, tables, charts, and test takers have 1 minute to prepare for the discussion task, which is the final task of the test. In the CET-SET Band 6, the higher level of the test, the communicative goal is to argue and debate on a topic of a somewhat abstract nature such as social issues, cultural differences, and so on. The prompts are mostly verbal, and there is no planning time for the discussion task. Following the discussion, test takers will answer a further-check question which is related to the discussion topic.

Table 4 Contextual features: the discussion tasks of the CET-SET Band 4 and Band 6

	CET-SET Band 4	CET-SET Band 6
Purpose of discussion	Reaching an agreement through co-operative discussion	Persuading the partner through argument and debate
Nature of information	Mainly personal, concrete	Mainly non-personal, abstract
Discourse mode	Exposition, description, comparison	Comparison, argument, persuasion

Table 4 (continued)

	CET-SET Band 4	CET-SET Band 6
Time constraints	Planning for 1 minute; discussing for 3 minutes	No planning time; discussing for 3 minutes
Task prompts	Picture, table, chart, verbal	Mainly verbal
Order of task	The final task in the test	The second part (total three parts)

Research and validation

Though performance conditions have been considered carefully to operationalise the constructs of interactional competence, the effects of these contextual variables on test takers' co-constructed performances need to be investigated.

Zhang (2004) looked into a number of factors that were likely to have an effect on test takers' performances on the group discussion task in the face-to-face CET-SET. The factors of interest included the topic of discussion, peer candidates' level of English proficiency, use of oral communication strategies, personality and anxiety of group members. Questionnaire surveys were conducted among CET-SET test takers and examiners, followed by a retrospective study for further evidence on the nature and extent of the impact on test takers' performances of the variables of interest. Topic familiarity and interestingness were found to have a significant impact on how much test takers had to say in the discussion. Proficiency of peer candidates and the use of communicative strategies by group members were also found to have affected test takers' performances in the discussion. The effect of peer candidates' personality and test anxiety on the interactivity of the group discussion was found to be relatively modest. The study drew the conclusion that 'the merits of co-construction of discourse, when used for testing one's communicative language ability, outweigh its limitations' (Zhang 2004:98).

He and Dai (2006) investigated the degree of interaction among candidates in the CET-SET group discussion. Using a 170,000-word corpus of test takers' performances on the discussion task, the study analysed the frequencies of occurrences of ILFs (see the section 'Operations: the discussion task'). The analysis revealed inadequate elicitation of six categories of ILFs from the candidates, which raised serious concerns over the validity of the discussion task. As the corpus was built using performances of test takers in one test centre, the source of the corpus may be seen as a limitation to the representativeness of the data. It is also likely that test takers were not well trained in skills and strategies for participating in interactive speaking tasks, given the data was collected in November 2001, two years after the CET-SET was

inaugurated. In any case, more research in this regard is needed to look into the interactive features of the discourses elicited in the discussion task after the test has been in operation for about two decades.

Jin and Zhang (2016) investigated the impact of test mode, face-to-face versus computer-based, on the use of communication strategies in the discussion task of the CET-SET. Through conversation analysis of test takers' performances on the discussion tasks, the study revealed a high level of similarities in the quantity and variety of communication strategies in the two discussion tasks. The study also showed that the test takers were generally capable of turn-taking effectively in the computer-based discussion task, though there seemed to be a neat and orderly turn-taking mechanism being co-constructed by most of the test takers. The findings also suggested that effective use of these strategies may help enhance test performance in a speaking task involving peer-to-peer interaction. It was noted that future studies need to consider the interaction among the test mode and other contextual variables (e.g. gender, personality, computer anxiety, and computer familiarity) and their interactive effects on test taker performances on the CET-SET discussion task.

Lessons learned from the two case studies

The most important lesson learned from the two cases is that careful specification and implementation of task operations and performance conditions are essential for operationalising assessment constructs. Only when assessment tasks are fully contextualised will assessment results be generalisable to communicative use of language in the real world. In this section, reflections on the two cases are detailed to highlight the role of contextual facets in operationalising assessment constructs.

Domain description for test operations

As mentioned previously, Weir (2013) suggested two key questions for consideration when the context validity of a language test is to be established. The first question is whether the characteristics of the test task constitute an adequate and comprehensive representation of those that would be normally encountered in the real-life context. The analysis of the two cases shows that the investigation and description of target language use situations constitute an essential first step towards ensuring the context validity of an assessment.

Khalifa and Weir (2009:81) noted that the starting point for the development of a reading assessment is to 'describe target reading activities in terms of their criterial parameters (context and cognitive)'. In the AERT project, a needs analysis was conducted to better understand students' purposes of reading, the types of texts they were likely to encounter, the strategies they

would use to process these various types of texts, and the status quo of English reading instruction and assessment in universities in China. Data was collected through a questionnaire survey and an analysis of EAP reading textbooks and assessment tasks. The results of the investigation facilitated the specification of contextually appropriate operations of EAP reading for Chinese learners of English at an advanced level. Following the needs analysis, the constructs of expeditious reading were defined to incorporate the contextual parameters of target language use, in this case, effective and efficient reading for the gist (skimming), main points (search reading) and specific details (scanning).

In the case of the CET-SET, the assessment incorporates a peer-to-peer discussion task to better represent the constructs of interactional competence. The decision on the group format in which three (or in rare cases four) test takers interact with each other was made on the basis of a national survey of the needs for oral communication in English prior to the development of the CET-SET (Huang 1999). In mainland China where English is learned and used as a foreign language, the educational domain is a relevant context of language use. The format of peer-to-peer discussion was therefore adopted to better represent language use in the educational context and provide test takers with authentic purposes of communication in task completion. Galaczi and ffrench (2011:121–122) pointed out that 'performance-based testing has witnessed an emphasis for assessment tasks to share features considered to be central in a classroom context'. It was further noted that the peer–peer interaction in the Cambridge English Speaking tests 'is in line with the general purpose of the tests to reflect classroom practices and the educational domain of language learning'.

Large-scale computer-based speaking tests such as the Internet-based Test of English as a Foreign Language (TOEFL iBT) and Pearson Test of Academic English (PTE Academic) typically employ monologic tasks only, due to the technical difficulty in delivering and scoring interactive speaking tasks in a computer-based format. When the face-to-face CET-SET was replaced by the computer-based format in 2013, efforts were made to retain the discussion task so that the context validity of the speaking test would not be compromised. The online interactive task in the CET-SET also has the advantage of simulating the increasingly popular form of computer-mediated oral communication such as talking via Skype or WeChat, participating in online discussion, and attending video conferences.

Specification of performance conditions

The second question Weir (2013) suggested for consideration is whether the characteristics of the test task are appropriately calibrated to the level of proficiency of the learner being evaluated. Domain description lends support

to the definition of test constructs in operational terms, making explicit the theoretical framework underlying the test. With an operational definition of the constructs, the next step is to establish the context validity by specifying performance conditions and fine-tuning contextual variables according to the level of proficiency of targeted test takers.

When the AERT expeditious reading tasks were developed, performance conditions were specified in great detail, including mainly text-based parameters (e.g. text length, text type, language difficulty, topic familiarity) and task-based parameters (e.g. purpose of reading, response method, weighting, timing). More importantly, evidence was collected to make sure that the conditions were met in task development. For example, to find out whether the texts pre-selected by item writers were suitable for assessing expeditious reading, questionnaire surveys were conducted among EAP teachers for their evaluation of the language difficulty, topic familiarity and subject specificity of the texts. After the pilot test of the prototype version, test takers were also surveyed for their perceptions of the suitability of the texts and tasks to assess their EAP reading abilities.

For the CET-SET, contextual variables of the discussion task were specified and adjusted to differentiate the two levels of proficiency: Band 4 and Band 6. Tasks at the lower level elicit co-operative discussions among the candidates on topics of a concrete nature. For example, in a CET-SET Band 4 discussion task, test takers are instructed to work out a travel plan by discussing places to visit, schedule of the visit and means of transportation. Test takers have 1 minute to prepare for the discussion. Tasks at the higher level, on the other hand, engage test takers in discussions on topics of a more abstract and controversial nature, and the topic of discussion in the CET-SET Band 6 is related to the theme of an individual presentation task, which precedes the discussion. For example, test takers make an individual presentation on the topic of retirement, with focuses on the ageing of the population and the current employment situation in China. Immediately following the presentations, they start the discussion on whether retirement age should be postponed. There is no preparation time for the discussion task at the higher level.

Challenges in establishing context validity

Important as it is to establish *a priori* validity for test development, achieving context validity is not without its problems due to 'the difficulty we have in characterising language proficiency with sufficient precision to ensure the validity of the representative sample we include in our tests, and the further threats to validity arising out of any attempts to operationalise real-life behaviours in a test' (Weir 2005:20).

The biggest challenge in achieving context validity is to have an in-depth understanding of the real-life communicative use of language and define

assessment constructs in operational terms. In other words, context validity needs to be established by *a priori* evidence on the extent to which tasks are representative of target language use situations. Although task operations of the AERT expeditious reading were developed in a principled way, types of the texts seemed to be limited by the sources available for assessment purposes (e.g. the copyright issue) and by the test time available for using more texts in each section. The text types featured in the AERT, therefore, may not sufficiently represent those to be encountered by English language learners at an advanced level in their future learning and working environments. Similarly, the peer-to-peer discussion task was adopted in the CET-SET to simulate language use in the educational domain. Interview by an oral examiner was considered problematic because the literature has suggested the issue of an imbalanced power relationship between the examiner and the candidate. Luoma (2004:35) pointed out that 'the fact that the interviewer has considerable power over the examinee in an interview has been recognised as one of the central weaknesses of this test type'. Nonetheless, engaging in discussions with highly proficient speakers (e.g. native speakers) is a relevant context of language use for university students, who are likely to participate in international conferences or collaborative projects with overseas partners, as well as interactive discussions with native speakers in future workplaces.

Construct operationalisation also presents a major challenge to the context validity of language assessments. With operational definitions of test constructs, tasks should be developed to simulate authentic use of language in real-life contexts. In the testing context, however, full authenticity of setting is not possible. Test methods inevitably contain construct-irrelevant factors that affect the accuracy of scores and the meaningfulness of score interpretations. Therefore, attempts should be made within the constraints of the test situation to approximate to 'situational authenticity' (Weir 2005:56). The implementation of the AERT expeditious reading tasks shows that no matter how well the tasks have been designed, test takers' cognitive processes may not be completely congruent with the expectations of the task designer. The failure to control the response time at the text level, for example, resulted in the contamination of the constructs of expeditious reading because inadequate time was spent on the final text in each section (Jin 2002, Weir et al 2000).

Bachman (1991:690) defined situational authenticity as 'the perceived relevance of the test method characteristics to the features of a specific target language use situation'. For a test task to be perceived as situationally authentic, Bachman explained, 'the characteristics of the test task need to be perceived as corresponding to the features of a target language use situation'. It can be seen that situational authenticity is essential to achieving context validity. To help test developers establish situational authenticity, Bachman and Palmer

(1996:57) developed a task characteristics checklist for comparing characteristics of target language use tasks and test tasks or creating completely new test task types.

However, even when situational authenticity is achieved, test constructs may not be truly operationalised. The corpus-based analysis of CET-SET test takers' ILFs revealed the discrepancy between the expected and the actual performances on the group discussion task (He and Dai 2006). It is understandable that in the testing context, test takers tend to take the opportunity to 'show off' their abilities, as indicated by their neat and orderly long turns in the CET-SET discussion task (Jin and Zhang 2016). Bachman (1991:691) proposed the concept of 'interactional authenticity', which was defined as 'a function of the extent and type of involvement of test takers' language ability in accomplishing a test task'. In contrast to situational authenticity, interactional authenticity resides in the interaction between the test taker and the test task, which is, in essence, what cognitive validity of the socio-cognitive framework is mainly concerned about. To achieve interactional authenticity, test takers should resort to relevant knowledge, skills and strategies and activate cognitive processes that resemble language use in the real world.

Conclusion

Weir (2005:56) highlighted the significance of specifying performance operations and conditions in a socio-cognitive approach to language testing (emphases in original):

> The last decade of the twentieth century saw a general decline in the prestige of psychometric, statistically-driven approaches to testing. In its place there has been a growing interest in the importance of *context*, in defining domain of use *performance conditions* and *operations*.

Context validity is no doubt an essential component of the socio-cognitive framework proposed in Weir (2005) and further developed and refined in O'Sullivan and Weir (2011). From a socio-cognitive perspective, contextual facets fulfil a bridging role between the constructs assessed in a language test and the communicative competence required in real-world situations.

In this chapter, the concept of context validity is revisited and the central role of contextually appropriate operations and conditions in construct operationalisation is illustrated through a detailed analysis of the cases of the AERT and the CET-SET. Weir and O'Sullivan (2017:87) reflected on the earlier Sino-British collaboration in language testing and commented that Weir's experiences in working on the AERT and the CET with the Chinese team have contributed to the development of the socio-cognitive theory:

> . . . the roots of the socio-cognitive framework arose out of this earlier collaborative work by Weir in China, first as senior UK consultant on the national College English Test . . . the framework was developed further in his consultancy work on the Advanced English Reading Test.

For language testers based in mainland China, working with Cyril Weir on the development and validation of language assessments in the 1990s involved developing a clearer specification of the operations and performance conditions underlying language test performance. The experiences have provided the conceptual basis for the contextual validity parameters specified in the socio-cognitive framework for validating language tests. Specifically, the experiences language testers gained through close collaboration with Cyril Weir have contributed significantly to a sustained development of the CET, a locally developed language test with distinctive Chinese characteristics which has continuously sought to meet the exacting professional standards set for high-stakes tests in the 21st century (Jin 2019).

References

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

Bachman, L F (1991) What does language testing have to offer?, *TESOL Quarterly* 25 (4), 671–704.

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

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

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

Chapelle, C A (1999) Validity in language assessment, *Annual Review of Applied Linguistics* 19, 254–272.

Davidson, F and Lynch, B K (2002) *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications*, New Haven: Yale University Press.

Galaczi, E and ffrench, A (2011) Context validity, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press, 112–170.

He, A W and Young, R F (Eds) (1998) *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency,* Amsterdam/Philadelphia: John Benjamins.

He, L and Dai, Y (2006) A corpus-based investigation into the validity of the CET-SET group discussion, *Language Testing* 23 (3), 370–401.

Huang, S H (1999) *The development and validation of the College English Test – Spoken English Test (CET-SET)*, unpublished PhD thesis, School of Foreign Languages, Shanghai Jiao Tong University.

Hymes, D (1972) Models of the interaction of language and social life, in Gumperez, J J and Hymes, D H (Eds) *Directions in Sociolinguistics*, New York: Holt, Rinehart and Winston, 35–71.

Jin, Y (2002) *The Development and Validation of the Advanced English Reading Test,* Shanghai: Shanghai Jiao Tong University Press.
Jin, Y (2010) The National College English Testing Committee of China, in Cheng, L and Curtis, A (Eds) *English Language Assessment and the Chinese Learner*, New York: Routledge, 44–59.
Jin, Y (2019) Testing tertiary-level English language learners: The College English Test in China, in Su, L I-W, Weir, C J and Wu, J R W (Eds) *English Language Proficiency Testing in Asia: A New Paradigm Bridging Global and Local Contexts*, London/New York: Routledge, 101–130.
Jin, Y and Zhang, L (2016) The impact of test mode on the use of communication strategies in paired discussion, in Yu, G and Jin, Y (Eds) *Assessing Chinese Learners of English: Language Constructs, Consequences and Conundrums*, Basingstoke: Palgrave Macmillan, 61–84.
Khalifa, H and Weir, C J (2009) *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing volume 29, Cambridge: UCLES/Cambridge University Press.
Luoma, S (2004) *Assessing Speaking,* Cambridge: Cambridge University Press.
McNamara, T (1997) Interaction in second language performance assessment: Whose performance?, *Applied Linguistics* 16 (2), 159–179.
Messick, S (1995) Standards of validity and the validity of standards in performance assessment, *Educational Measurement: Issues and Practice* 14, 5–8.
National College English Testing Committee (1999) *CET–SET Syllabus,* Shanghai: Shanghai Foreign Language Education Press.
National College English Testing Committee (2016) *Test Syllabus of the College English Test (Revised in 2016),* Shanghai: Shanghai Jiao Tong University Press.
O'Sullivan, B (2002) Learner acquaintanceship and oral proficiency test pair-task performance, *Language Testing* 19 (3), 277–295.
O'Sullivan, B and Weir, C J (2011) Language testing and validation, in O'Sullivan, B (Ed) *Language Testing: Theory and Practices*, Basingstoke: Palgrave Macmillan, 13–32.
Taylor, L (Ed) (2011) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press.
Taylor, L and Galaczi, E (2011) Scoring validity, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press, 171–233.
Taylor, L and Wigglesworth, G (2009) Editorial: Are two heads better than one? Pair work in L2 assessment contexts, *Language Testing* 26 (3), 325–339.
Urquhart, A and Weir, C J (1998) *Reading in a Second Language: Process, Product and Practice*, Harlow: Pearson Education Ltd.
Weir, C J (1990) *Communicative Language Testing*, New York: Prentice Hall.
Weir, C J (1993) *Understanding and Developing Language Tests*, New York: Prentice Hall.
Weir, C J (2005) *Language Testing and Validation: An Evidence-based Approach,* Basingstoke: Palgrave Macmillan.
Weir, C J (2013) An overview of the influences on English language testing in the United Kingdom 1913–2012, in Weir, C J, Vidaković, I and Galaczi, E D, *Measured Constructs: A History of Cambridge English Language Examinations 1913–2012,* Studies in Language Testing volume 37, Cambridge: UCLES/Cambridge University Press, 1–102.

Weir, C J and O'Sullivan, B (2017) *Assessing English on the Global Stage: The British Council and English Language Testing 1941–2016*, Sheffield: Equinox.

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

Weir, C J, Yang, H and Jin, Y (2000) *An Empirical Investigation of the Componentiality of L2 Reading in English for Academic Purposes,* Studies in Language Testing volume 12, Cambridge: UCLES/Cambridge University Press.

Yang, H and Weir, C J (1998) *The Validation Study of the College English Test,* Shanghai: Shanghai Foreign Language Education Press.

Young, R F (2008) *Language and Interaction: An Advanced Resource Book*, London/New York: Routledge.

Zhang, L (2004) *Merits and Limitations of Co-construction of Discourse in an Oral Interview Test: An Analysis of Major Factors That Might Affect Candidate Performance in Group Discussion,* unpublished MA thesis, School of Foreign Languages, Shanghai Jiao Tong University.

Zheng, Y and Cheng, L (2008) The College English Test (CET) in China, *Language Testing* 25 (3), 408–417.

5 Placing construct definition at the heart of assessment: research, design and a priori validation

Sathena Chan

Centre for Research in English Language Learning and Assessment, Bedfordshire

Nicola Latimer

Centre for Research in English Language Learning and Assessment, Bedfordshire

In this chapter, we will first highlight Professor Cyril Weir's major research into the nature of academic reading. Using one of his test development projects as an example, we will describe how the construct of academic reading was operationalised in the local context of a British university by theoretical construct definition together with empirical analyses of students' reading patterns on the test through eye-tracking. As we progress through the chapter we reflect on how Weir's various research projects fed into the development of the test and a new method of analysing eye-tracking data in relation to different types of reading.

Introduction

One of the key themes in Weir's work was that of test validity. Time and time again Weir advocated that a key component (if not the starting point) of validity was construct definition. This relies upon tests eliciting the core cognitive processes as would be demanded of test takers in the world beyond the test, a view shared by Davies (1984:50–69): 'in the end no empirical study can improve a test's validity . . . What is most important is the preliminary thinking and the preliminary analysis as to the nature of the language learning we aim to capture.'

To achieve this, Weir insisted on the importance of gaining a thorough understanding of what it was that students had to do when they read at university. Understanding this enabled Weir to develop his model of reading – a model which illustrates that students operationalise different types of reading in response to their purposes for reading. Weir's influential model

of reading (Khalifa and Weir 2009, Weir 2005) has been widely used by test developers in the UK and worldwide (including British Council, Cambridge Assessment English, Language Training and Testing Center (LTTC) in Taiwan and EIKEN in Japan) to define the construct of reading at different proficiency levels in different contexts (e.g. educational, academic and professional). In his final presentation at the CRELLA Spring Research Seminar (March 2018), he called for reflection to evaluate the extent to which current academic reading tests are fit for purpose.

Weir's research into the nature of reading (reading skills and reading texts)

Weir conducted research into the nature of academic reading over 35 years. In his PhD research, Weir (1983) investigated the language activities and associated problems of students studying at tertiary level throughout the UK. Data was collected from over 2,000 overseas students, British students and academic staff. Weir, Yang and Jin (2000) conducted a similar analysis in an English as a foreign language (EFL) context. They investigated the academic English needs of first-year undergraduates in China. The results showed that students were expected to employ different types of reading in response to various academic reading needs. Nevertheless, due to lack of awareness and training, many students relied heavily on local careful reading (i.e. to read every sentence carefully and slowly) on all occasions. The different types of reading are explained in more detail in the next section.

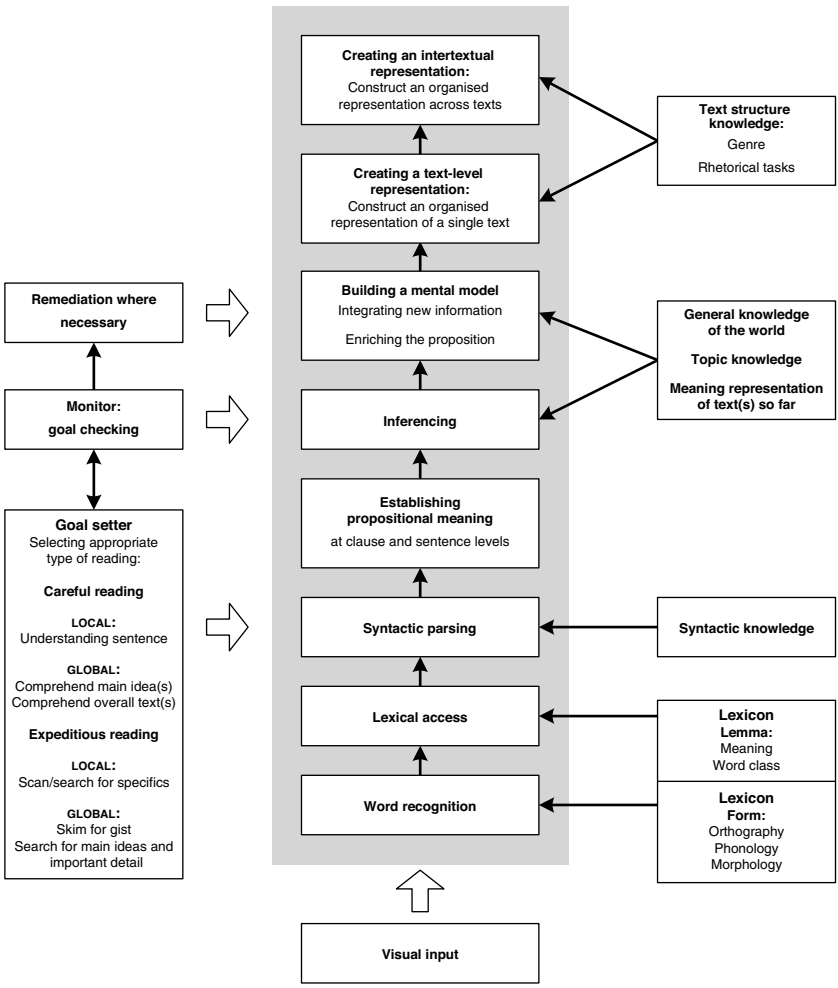
To define academic reading, another line of Weir's investigations concerned the properties of the reading texts themselves. Weir, Hawkey, Green, Ünaldi and Devi (2012) investigated the features of academic texts which most students from 14 subject areas encountered in their studies. For the first time, the level of text complexity required at undergraduate courses was established in relation to a range of textual measures such as average sentence length, syntactic complexity, and complexity of vocabulary. The explicit profiling of academic texts made it possible for test developers to align the reading texts in their tests to these measures, so as to ensure they reflect the same level of difficulty as texts that students will encounter on their courses.

Weir's model of reading: types of reading guided by the reader's goal

Understanding what was demanded of students helped Weir to develop a model of reading (Weir 2005, Khalifa and Weir 2009). Based on a componential view of reading (a view which considers reading as a set of sub-skills, see Weir et al 2000), Weir's model illustrates that readers operationalise different *types of reading* in response to their *reading purposes*. This 'goal setter'

process (in the left-hand column in Figure 1) determines how the reader will engage with the text, deciding what types of reading to employ. As the result of the goal setter, the reader makes critical decisions which affect the level(s) of processing to be activated from word recognition through to creating an intertextual representation (see the central core of the model). Skilled readers would normally 'monitor' how well they read in response to their goals. The knowledge base required for comprehension constitutes the right-hand column. These latter features reappear under the contextual parameters discussed below which determine the cognitive load of the text to be processed.

Figure 1 Weir's reading model (Khalifa and Weir 2009:43, Weir 2005)



A significant contribution of Weir's reading model is that it explicitly specifies different types of reading and levels of cognitive processes in reading. High-level processes tend to be conscious and effortful whilst low-level processes are automated and largely subconscious for most skilled readers. Although it is now generally accepted that readers process at different levels simultaneously in order to establish meaning in reading, the explicit description of the different types of reading and levels of reading processes in the model has helped researchers to evaluate the cognitive demand of reading tasks against the target processes and to develop new reading tests.

Weir argued that any reading test should distinguish between careful and expeditious reading at local and global level. Khalifa and Weir (2009:46) suggest that *careful reading* 'is intended to extract complete meanings from the presented material at a local or a global level, i.e. within or beyond the sentence right up to the level of the complete text or texts', a view that accords with the type of careful reading described by Rayner, Pollatsek, Ashby and Clifton (2012). This type of reading is based on slow, careful, linear, and incremental reading for comprehension. *Expeditious reading*, on the other hand, includes skimming, search reading, and scanning (Urquhart and Weir 1998). *Skimming* or gist reading is generally defined as reading to obtain the gist, general impression and/or superordinate main idea of a text. It takes place when the reader attempts to build a broad understanding (a macrostructure) of the text by reading very selectively, reading the minimum amount of information possible. Skimming is necessarily a form of global reading as it must encompass several ideas or propositions distributed across the wider text. *Search reading* involves locating information on predetermined topics. The reader only wants the information necessary to answer set questions or to extract data, for example in order to complete written assignments. Search reading differs from skimming in that the search for information is guided by predetermined topics so the reader does not necessarily have to establish a macro-propositional structure for the whole of the text. Search reading can take place at both the local and global level. Where the desired information can be located within a single sentence it would be classified as local and where information has to be put together across sentences it would be seen as global. In both cases the search is for words in the same semantic field as the target information, unlike *scanning* where exact word matches are sought. Khalifa and Weir (2009) argued that scanning should always be considered local. This is not because the scan for the word is confined to a single sentence but because the item sought (a single word or phrase) operates at a local level. Studies into students' reading abilities have indicated that for many readers, reading quickly, selectively and efficiently poses greater problems than reading carefully and efficiently (Beard 1972, Weir 1983, Weir et al 2000) because it demands rapid recognition which is contingent upon sufficient practice in reading in the target language.

In summarising Weir's notion of academic reading, the following points can be highlighted:

- when reading for academic purposes, several types of reading are likely to emerge
- it is important to teach/assess both careful and expeditious global reading
- there is a need for students to process and integrate information at both the whole text and intertextual levels, rather than just comprehending at the paragraph or sentence levels
- texts used in reading tests should mirror the contextual variables found in real-life academic texts
- separability of reading skills should be mirrored in the way results are reported.

The reader is referred to Weir and Chan (2019) for a synthesis of research on assessing academic reading.

Personal reflections: assessing careful and expeditious reading skills

Cyril's model of reading, developed over a number of years, exemplified his concern with ensuring that theory related to the real world of language use. Previous models of reading sought to provide a detailed model of the reading process (e.g. Stanovich 1980) or identify the components which contribute to reading ability (e.g. Hoover and Tunmer 1993) but did not account for how readers deploy reading in different ways for different purposes. For Cyril, without this, their explanation of how reading operates was incomplete. Some of his early work reflected a more cautious approach to a componential approach (Weir and Porter 1996), suggesting that more research was necessary. But his work developing the Advanced English Reading Test (AERT) (Weir et al 2000) for undergraduates in China illustrates that by then Cyril was convinced that a componential approach was necessary. AERT included five different sections each aimed at a different type of reading (careful global, expeditious global: a skimming task and a search task, expeditious local: a scanning task, and careful local). Weir et al (2000:23) explained the rationale behind this design, writing that 'reading is at the very least a bi-divisible process. For the benefits of teaching and testing, a unitary view of reading should be discarded'. Subsequently, Cyril became a strong advocate for assessing both careful and expeditious forms of reading in preparation for the academic world. Other English for Academic Purposes (EAP) reading tests Cyril developed such as Test of English for Academic

Purposes (TEAP) (Eiken Foundation of Japan and Sophia University) and GEPT Advanced (The General English Proficiency Test) (LTTC, Taiwan) have dedicated sections on different types of reading. This approach to assessing academic reading was again clearly visible in the design of the Bedfordshire Academic Reading Test (BART).

Operationalisation of the academic reading construct in a local context

Throughout his career, Weir developed several university-based academic reading tests. A good example is the Test in English for Educational Purposes (TEEP) (formerly known as AEB/TEAP) he developed at the University of Reading in the 1980s. The test is still in use today. During his tenure as the Powdrill Professor in English Language Acquisition at the University of Bedfordshire (2005–18), he developed the Bedfordshire Academic Reading Test (BART). Using BART as an example, we will describe how the construct of academic reading was operationalised in the local context of a British university.

Aim of the test

Weir sought to provide universities with a quick, cost-effective and robust method for identifying new students that are likely to need support with their academic reading and writing skills. Therefore, BART was developed to be taken by all (home and international) students after entry to the university to support students' learning, regardless of whether English is their first language. The test is diagnostic in nature, offering both an indication of the extent of support required as well as providing insight into any areas of concern. The test has a reading paper and a reading-into-writing paper. The key stages of test development are provided in Appendix 1.

Structure

The reading paper is designed to reflect the types of reading identified by Weir's model of reading. Part 1 is concerned with careful reading; students are expected to identify Text 1's main ideas. Part 2 is concerned with both careful and expeditious reading as they relate to creating an intertextual representation. Part 2 requires students to read seven short mini-texts carefully, and then use expeditious reading skills to search through the text in Part 1 for semantic links and match the paragraphs which share the same theme as the mini-texts. Part 3 targets the skills and strategies required to process large

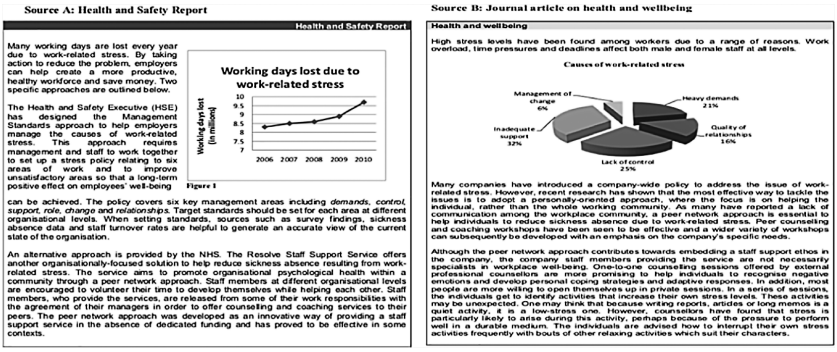
amounts of extended text selectively and efficiently. Students are expected to use expeditious reading skills to read Text 2 and select a heading for each paragraph (see Table 1).

Table 1 Structure of reading paper

Part	Text	Task	Time
1	Text 1 (approx. 1,000 words)	Select a summary statement for each paragraph which accurately summarises the main point made in that paragraph. There is a list of 10 statements to choose from.	20 minutes
2	Seven short mini-texts (approx. 70 words each) which share the same theme as Text 1	Match each mini-text to the paragraph from Text 1.	20 minutes
3	Text 2 (approx. 1,000 words)	Select an appropriate heading for each paragraph from a list of eight headings.	10 minutes

The reading-into-writing paper aims to test students' academic literacy skills when the subject matter and information to be communicated by the writing must be sourced from reading materials, as in an academic assignment or essay. Students are required to write an essay of 200–250 words in 60 minutes, summarising the main propositions from two one-page articles (each with a non-verbal input, e.g. a chart, a table or a diagram) and providing recommendations on the issue with a personal interpretation (see Figure 2 on page 112). Typically, students receive a test booklet (which includes the task instruction and two articles) on paper but they are required to compose the essay on a computer in a Word document. The two articles are on the same topic, e.g. work-related stress. The two articles describe the issue and suggest several solutions to reduce stress in the workplace. However, it is designed that the two articles share a few propositions, e.g. one solution is mentioned in both articles. To score well the students need to include the most relevant information from both sources and organise their writing in response to the task. Students are not penalised for poor referencing practice; however, they are warned against copying chunks of the text.

Figure 2 An example of the reading input for the reading-into-writing task



Scoring approach

The reading paper score is a composite of scores for three components, i.e. ability to read carefully to develop a text-level representation, ability to connect information across texts, and ability to read quickly and selectively for main ideas. Similarly, the reading-into-writing paper adopts a mark scheme whereby the total score is accumulated from scores in three analytic categories which represent the key construct of academic reading-into-writing skills (Knoch and Sitajalabhorn 2013), i.e. relevance and adequacy of content, organisation and language. Students' performances can be rated either 3, 2 or 1 on each category.

Students receive a profile of their academic literacy skills in relation to the three reading skills and three reading-into-writing criteria. Based on their total scores (on both the reading and reading-into-writing papers), students are categorised into three groups, according to a type of 'traffic light' system: *Needs comprehensive support (red)*/*Needs some support (amber)*/*Needs no support (green)* in academic literacy skills. Students who are in the red group are recommended (though not required) to take an intensive course on academic literacy provided by the university. Students who are in the amber group are seen by an academic literacy advisor to discuss a self-learning plan to improve their academic literacy skills. Students who are in the green category could make an appointment with an academic literacy advisor to discuss any issues they might have.

Personal reflections: his passion for promoting academic literacy

It may seem obvious to say that universities need to be able to identify those students with a deficit in academic skills at an early stage, and this rather begs the question as to why all universities do not do this as a matter of course. The reality is that this is extremely difficult to execute in a practical, timely, cost-effective and reliable manner. During the development of BART, Cyril was hands-on at all levels: from securing funding, communicating with top management, lecturers and students, developing the test from test specifications through to writing items. He respected all aspects of his work and he always held himself to the highest standards. There were moments where there was no support to sustain the project. But Cyril just simply wouldn't give up on what he believed was beneficial for the students and the university. His persistence was the sole reason the project survived. The project reflected Cyril's passion for developing practical solutions to assessment problems underpinned by empirical research and sound theory. The test specifications and sample items underwent many rounds of revisions. His eyes lit up every time we found something to improve. We were also struck by how much he actually enjoyed writing the test items!

Personal reflections: importance of test validation

Cyril's socio-cognitive framework, first elaborated in Weir (2005) placed great emphasis on ensuring that test tasks require candidates to engage in the same cognitive processes during the test as will be demanded of them in life beyond the test. Cyril's passion for test validation and measuring a test's usefulness against the demands faced by candidates in life after the test can perhaps be traced back to the start of Cyril's own journey in language testing and assessment. Cyril's PhD thesis (Weir 1983) examined the language problems faced by overseas students studying in higher education in the UK. His PhD work not only surveyed the problems that non-native speaking students encountered in their academic lives but used this information to construct a framework of requirements for an academic language test. When you read the concluding chapter of his thesis, it serves to underline that Cyril was his own fiercest critic. His exacting standards and insistence on empirical investigation are made clear.

Use of eye-tracking in test validation

Another contribution of Weir's (2005) socio-cognitive framework is that it urges the need for *a priori* test validation. It is important to collect evidence to show the extent to which the test (or items) elicits the target language skills as intended before test launch. Various methodologies such as think-aloud protocols (Bråten and Strømsø 2003) and questionnaires (Chan 2018a, 2018b, Wu 2014, Yu 2005) have been used to examine students' reading-into-writing processes. Nevertheless, these methods largely depend on a subsequent recollection of reading activities.

With advances in technology, eye-tracking now offers an alternative method to record, in detail, students' eye movements as they read. Starr and Rayner (2001:156) suggest that '(f)or the most part, eye-movement data have proved to be highly reliable and useful in inferring the moment-to-moment processing of individual words and larger segments of text'. However, technologies as such provide an enormous amount of data in detail, e.g. individual fixations in milliseconds. One can sympathise with the challenge faced by researchers when analysing eye-tracking data. Common eye-tracking measures include the number of fixations, total fixation duration, mean fixation durations, saccades and regressions, see for example Bax (2013), Bax and Chan (2019) and Brunfaut and McCray (2015). These measures, which indicate when and where fixations occur, to some extent, allow researchers to compare the temporal features of reading across students under test conditions. Nevertheless, the relationship between the eye-tracking measures and types of reading a test aims to measure is largely an under-researched area. Whilst measures in isolation may be of limited value, differences in patterns formed by various eye-tracking measures offer an opportunity for better insight in relation to students' use of different types of reading. The validation study, therefore, aimed to answer the following research questions (RQs).

Research questions

1. What reading patterns emerged from the eye-tracking data?
2. To what extent did the BART reading-into-writing test elicit different types of reading?

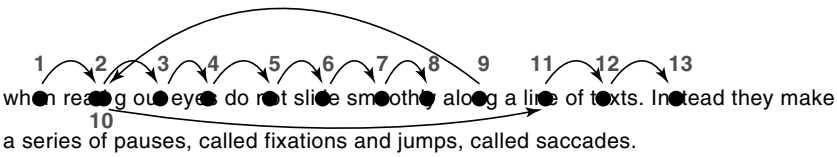
Eye-movement measures

When we read, although we may feel that our eyes slide along the line, our eyes actually make a series of jumps (**saccades**) separated by short periods when our gaze remains fixed on a word/part of a word (**fixations**). Rayner et al (2012) report that fixations typically only last about 250 milliseconds (a quarter of a second) with saccades taking even less time (typically about

40 milliseconds). Figure 3 illustrates a series of fixations and saccades as a reader progresses through a sentence.

In Figure 3, we can see that the reader fixates on 'when' (fixation 1), 'reading' (fixation 2), 'our' (fixation 3), 'eyes' (fixation 4) before skipping the word 'do' and fixating on the word 'not' (fixation 5). Some longer words such as 'smoothly' attract more than a single fixation (fixations 7 and 8). It is also evident that the eyes do not progress through the text in an entirely systematic way. When the reader reaches the word 'along' (fixation 9) there is a movement back to an earlier part of the sentence ('reading', fixation 10) before returning to the former location (fixation 11) to resume progress through the text. These backward glances are called **regressions**. Based on a study of undergraduates' reading patterns, Rayner et al (2012) reported that about 10–15% of fixations are regressions when adults read materials such as college textbooks. Generally speaking, words which are more familiar to the reader (high-frequency words) are likely to attract shorter fixations than uncommon words (low-frequency words) (Just and Carpenter 1980, Rayner 1977). Words which are more predictable from the context or the preceding text are also likely to attract shorter fixations (Ehrlich and Rayner 1981, Zola 1984). As texts become more challenging (grammatically more complex, less familiar content) fixation duration is likely to increase, saccade length to reduce and the percentage of regressions to increase (Rayner et al 2012:96). Although individual differences between students result in variations in fixations and

Figure 3 Illustration of a series of fixations and saccades



Personal reflections: what does it mean?

Cyril enjoyed academic discussions – one of the questions he often asked was 'what does it mean?'. This simple question pushed us beyond our comfort zone. In the context of this study, it is comparatively straightforward to report individual measures of students' eye movements on a test. But what can individual measures of eye movement tell us about their reading skills? It is, therefore, our hope to establish a method of analysing eye-tracking data which moves beyond the limits of individual measures to interpreting patterns of eye movement in relation to types of reading.

saccades, Fisher (1983) argues that changes in students' patterns of reading are observable across tasks.

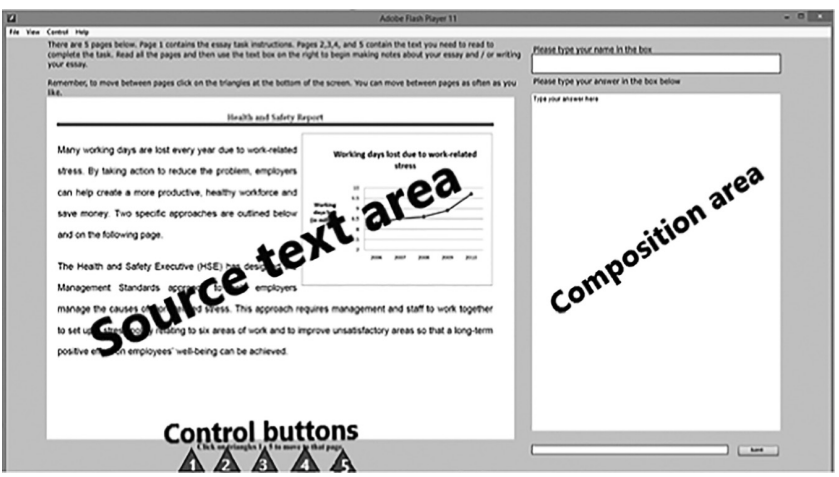
Participants

The eye-tracking study examined the eye movements of 30 C1 participants as they completed the BART reading-into-writing task. Students were recruited from several UK universities. Students were asked to self-rate their English proficiency using a set of Can Do statements (covering speaking, listening, reading and writing skills). Only students who rated themselves as C1 in four skills were included. 15 of the participants were first-year undergraduates and 15 were final-year undergraduates or postgraduates. 25 of the participants were native English speakers and five were non-native speakers. There was no significant difference in mean test scores between the native and non-native participants.

Task

One version of the BART reading-into-writing task was used, see Figure 2. In order to eye-track the participants as they worked, it was necessary to digitise the task so that it could be presented to participants on an interactive website (henceforth referred to as the *interface*). The interface (illustrated in Figure 4) included two main compartments: the source text area and the composition area. There were five pages of information within the source text area: one page of task instructions and four pages of source texts. Control buttons

Figure 4 A screen shot of the task interface

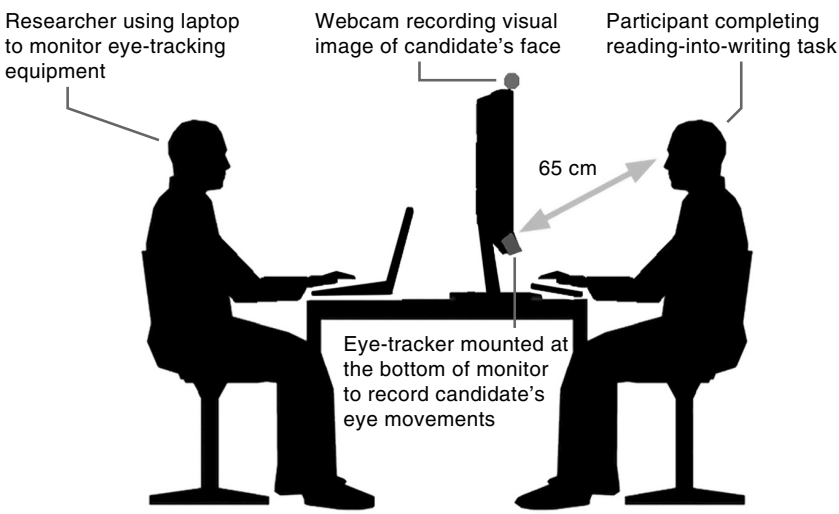


(i.e. triangles numbered 1–5 at the bottom of the screen) allowed the participants to 'turn the page' to select the materials they wished to read alongside their composition. Only the content in the source text area changed when the page was turned. The rest of the screen, including the composition area where participants typed their answer, remained unaffected.

Eye-tracker

A Tobii X2-60 eye-tracker was used to collect the eye-tracking data. Tobii (2018) reports accuracy of 0.4–0.6 of a visual degree and precision of 0.34–0.74 of a visual degree at a distance of 450mm–800mm. The range of participants' movements towards or away from the monitor, while seated in front of it, fell within these distances. Using an average of 0.5 of a degree for accuracy, the X2-60 was accurate to within 6mm on the screen at a range of 450mm–800mm. Figure 5 illustrates the setup of the eye-tracking equipment. When the source texts were displayed on the screen, characters were, on average, 7mm (or 25 pixels) high with a clear 12mm (or 43 pixels) between lines of text. Hence, with the eye-tracker accurate to 6mm (21 pixels) and a gap of 12mm between lines, it was possible to establish clearly which line of text was being fixated and, within a letter or two, which word on the line was being fixated.

Figure 5 Arrangement of the eye-tracking equipment for the data-collection sessions



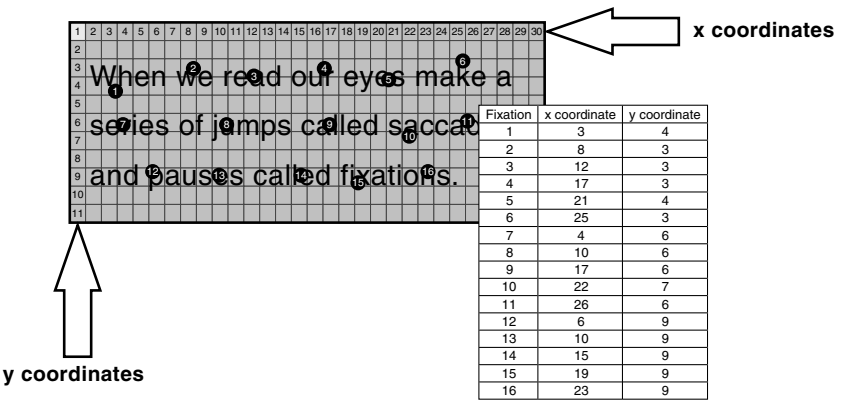
Data collection

Data was collected on a one-to-one basis. Each participant completed the reading-into-writing task on the interface (a maximum of 60 minutes was permitted) while their eye movements were recorded by the Tobii X2-60. After participants completed the task, semi-structured interviews, taking approximately 15 minutes, were conducted. The session ended with the participants completing a short cognitive processing questionnaire. Due to the focus of this chapter, we will only report the eye-tracking data.

Analysing the eye-tracking data

The participants' eye movements during the BART reading-into-writing task were recorded in terms of the exact point on the screen where participants focused. The resulting data offered a fixation-by-fixation account of where each participant had focused as they proceeded through the task. Figure 6 offers an illustration of the principle.

Figure 6 Illustration of how fixations are represented by coordinate data



Fixation	x coordinate	y coordinate
1	3	4
2	8	3
3	12	3
4	17	3
5	21	4
6	25	3
7	4	6
8	10	6
9	17	6
10	22	7
11	26	6
12	6	9
13	10	9
14	15	9
15	19	9
16	23	9

Due to the focus of the chapter, we only discuss results concerning how participants engaged with the written content of the source texts (i.e. the fixations on the source text area, excluding diagrams/visuals, in Figure 4). Each fixation was coded according to the area of the screen (*broad area of interest*) in which it occurred as well as the sentence on which the fixation had occurred (*sentence-level area of interest*). Using an algorithm, the data was then analysed and coded according to patterns that suggested the type of reading used as participants completed the task. It is beyond the scope of this chapter to provide a comprehensive explanation of the algorithm; for more details see Latimer (2018). In short, the coding algorithm first divided the fixation data

into *episodes* of reading. Episodes of reading started or ended when a break in reading had occurred (for example, if the participant had looked away from the screen for more than a second or moved their gaze to another area of the screen or changed page). Within each episode of reading, fixations were assigned to a category of short forward (SF), long forward (LF), short regression (SR) or long regression (LR) according to the properties outlined in Table 2.

Table 2 The classification of fixations according to the direction and distance moved through the text when compared to the previous fixation

Short forward (SF)	Any fixation which progresses **fewer** than 16 character spaces forward through the text (including moving from the end of one line to the beginning of the line below, which is termed a ***return sweep***).
Long forward (LF)	Any fixation which progresses **more** than 16 character spaces forward through the text (this includes any jump of more than 16 characters on the same line or any jump from one line to any line below **excluding *return sweeps***).
Short regression (SR)	Any fixation back to an earlier point on the same line or to the line above.
Long regression (LR)	Any fixation back to two or more lines above.

Once each fixation within an episode of reading was classified, the data was screened for patterns formed by the classifications. In order to deduce use of different types of reading, three criteria were used, i.e. distance and direction between fixations, proportion of regressions, and sentence boundary. The results of the coding by reading types are discussed in the next section.

Findings: evidence of test takers' reading skills from a priori test validation

Overall measures

In total, the 30 participants generated 215,052 fixations on the reading-into-writing task. The mean number of fixations per participant was 7,168 (SD: 1,839). These fixations added up to over 13 hours of data (about 27 minutes per participant). Of these fixations, about 30% were on the written source texts.

Table 3 Fixation data on the written source texts

	Mean	**SD**
Number of fixations per participant	2,646	872
Total fixation duration (minutes:seconds) **per participant**	8:07	2:46
Fixation duration (milliseconds) **per fixation**	184	100

As shown in Table 3, participants overall spent about eight minutes on the source texts (excluding the diagrams) and had an average fixation duration of 184 milliseconds (i.e. less than a fifth of a second). It is interesting to note that the average fixation duration on the source texts was shorter than the mean fixations reported in the literature for reading. For example, Rayner et al (2012) reported mean fixation durations of around 200 milliseconds for light fiction, through to a mean of 260 milliseconds for more complex scientific texts. Brunfaut and McCray (2015) reported a mean fixation duration on short B2-level passages of 237 milliseconds. The mean of 184 milliseconds in this study is comparatively short. In addition, this study recorded a noticeably higher regression percentage (34%) than that reported in the Rayner et al study (11%) or the Brunfaut and McCray study (19% for B2-level text). High rates of regression could indicate readers experiencing comprehension difficulties or their use of selective reading (see further discussion in the subsequent 'Selective reading' section).

RQ1: Patterns of reading that emerged from eye-tracking data

Based on the eye-tracking data, we were able to distinguish four types of reading: careful local reading, careful global reading, selective local reading and selective global reading. Each will be described in the next sections, followed by a discussion of how these compare to the categorisation in Khalifa and Weir's (2009) model.

Careful reading

Careful reading was relatively straightforward to establish from the eye-tracking data. Reading episodes were coded as **careful reading** when they formed part of a pattern which progressed methodically through the text in a linear pattern. Careful reading fixations:

- moved forward through the text with short forward-moving saccades of no more than 16 character spaces (the average careful reading saccade being eight characters, Rayner et al 2012)
- did not skip over any areas of text
- regressed to earlier parts of the text at a rate of less than one regression to every three forward-moving fixations.

If the episode of careful reading remained within a single sentence, it was regarded as **careful local reading**. Once the episode of careful reading extended beyond a single sentence, it became **careful global reading**.

Selective reading

Exceeding the parameters for careful reading, either in terms of distances between fixations, skipping over parts of the text or in terms of the rate of regressions, resulted in fixations being coded as **selective reading**. It is important to note that we had difficulty mapping this category of selective reading observed by eye-tracking data to the expeditious reading described in Weir's reading model. As described earlier, Weir's expeditious reading can operate in the forms of skimming, search reading and scanning. Although Khalifa and Weir (2009) used the same terminology as Urquhart and Weir's matrix of reading types (1998:123), there are some differences relating to search reading. Urquhart and Weir suggested that scanning should always be considered local whilst skimming and searching always operate at a global level. Khalifa and Weir (2009) agreed that scanning should always be considered local because the item sought (a single word or phrase) operates at a local level and skimming should always be considered global (because the reader is attempting to build a broad understanding of the whole text). However, Khalifa and Weir (2009) argued that search reading can indeed operate at both a local and a global level. In other words, for both Urquhart and Weir (1998) and Khalifa and Weir (2009), the global and local classifications concern primarily whether the reading activities resulted in comprehension at a local level (i.e. a single word or idea) or a global understanding of the whole text. On the other hand, the global and local classifications of the algorithm in the current study relate to the actual locations of reading, i.e. whether the episode of reading extended beyond a single sentence (global) or remained within a single sentence (local).

As a result, it was decided that in this study, reading patterns which did not meet the linear, incremental pattern of careful reading were termed selective reading rather than expeditious reading to emphasise that they did not neatly align to Weir's categories. Nevertheless, allowing for the fact that the categories of selective reading in this study cannot be used directly to infer occurrences of scanning, skimming and search reading, the method, for the first time, allows researchers to differentiate between students' eye-tracking evidence of careful reading and selective reading on a test (see Table 4 for a summary).

Table 4 Patterns of reading that emerged from eye-tracking data in relation to Weir's models (adapted from Latimer 2018)

Algorithm classification of eye-tracking data	Fixation patterns	Urquhart and Weir's matrix of reading types (1998)	Khalifa and Weir's model of reading (2009)
Careful local	When reader progresses through a single sentence in a linear, incremental pattern with fixations spaced no more than 16 characters apart and with regressions accounting for less than 25% of fixations.	**Careful local reading:** to gain a full understanding at sentence level.	
Careful global	When reader progresses through the text (extending beyond a single sentence) in a linear, incremental pattern with fixations spaced no more than 16 characters apart and with regressions accounting for less than 25% of fixations.	**Careful global reading:** to gain a full understanding at paragraph/text levels.	
Selective local	When reader has made a series of fixations which remain within a single sentence, but which do not progress through the sentence in a linear, incremental way (fixations are spaced more than 16 characters apart or more than 25% of fixations are regressions).	**Expeditious local: Scanning** (possibly across the whole text) to locate a specific word or piece of information.	**Expeditious local: Scanning** (possibly across the whole text) to locate a specific word or fact. **Searching** for information on a predetermined topic which can be obtained from a single sentence.
Selective global	When reader has made a series of fixations which extend beyond a single sentence, but which do not progress through the text in a linear, incremental way (fixations are widely spaced, more than 16 characters apart, or more than 25% of fixations are regressions).	**Expeditious global: Skimming** to establish topic and main ideas. **Searching** to locate and understand information.	**Expeditious global: Skimming** to establish the gist or main idea. **Searching** for information on a predetermined topic where the information needs to be put together across sentences.

RQ2: Test takers' use of different reading patterns

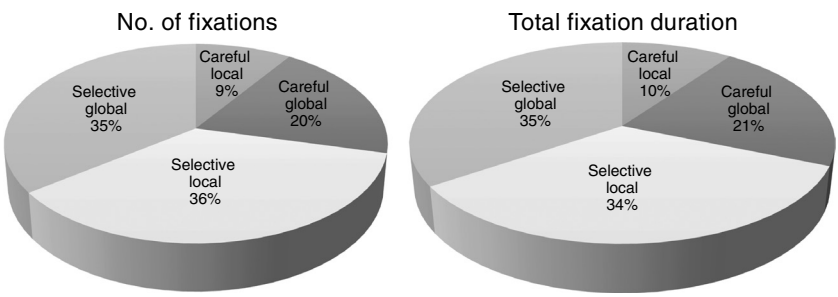
After reporting the reading patterns that emerged from the eye-tracking data, we now report the extent to which the participants showed each reading pattern (see Table 5). To remind the reader, the participants fixated on average a total of 27 minutes on screen during task completion. 30% of the total fixation on screen (i.e. about 8 minutes per participant) was spent on reading the two passages (539 words in total). As shown in Table 5, participants in this study displayed all four reading patterns. Nevertheless, it is interesting to note that they appeared to engage in selective reading patterns (i.e. non-linear, skipping fixations and/or regressions) more than careful reading patterns (i.e. linear, steady progression of fixations).

Table 5 Reading patterns on source texts

Reading patterns	No. of fixations of all participants	Total duration of all participants (hh:mm:ss)	No. of fixations per participant	Total duration per participant (hh:mm:ss)
Careful local	7,459	00:24:12	248.63	0:00:48
Careful global	15,515	00:51:58	517.17	0:01:44
Selective local	28,421	01:23:26	947.37	0:02:47
Selective global	27,989	01:23:58	932.97	0:02:48
Total	79,384	04:03:34		

As illustrated in Figure 7, careful reading patterns accounted for only 30% of the reading activity on the source texts. Careful global fixations accounted for 20% of total number of fixations and 21% of total duration. Careful local fixations accounted for around 10% of both total number and total duration. In contrast, selective reading patterns accounted for over 70% of the reading activity. Selective local fixations accounted for over a third (36% of number

Figure 7 Distribution of attention on written source texts according to reading type



and 34% of total duration) of the fixations. Selective global fixations were very similar, accounting for 35% of number of fixations and 35% of total fixation duration. The results, at the very least, show that participants altered their eye-movement patterns during the completion of the task. This provides some support to the theory that the task, as intended, required use of different types of reading.

Reading patterns of year-one undergraduates (Y1) and year-three undergraduates/postgraduates (Y3+)

The data for participants with more academic experience was compared to the data for participants with less academic experience. When the overall fixation data was compared between the two groups (Y1 and Y3+), the total number of fixations and the total fixation duration were very similar, see Table 6. Participants in both groups had an average of about 2,600 fixations on the source texts for about 8 minutes.

Table 6 Comparison of fixations on source texts between Y1 and Y3+ participants

Attention on written source texts	Total number of fixations	Total fixation duration (hh:mm:ss)	Mean total number of fixations per participant	Mean total fixation duration per participant (hh:mm:ss)
Y1 (n=15)	40,362	02:01:37	2,690.80	0:08:06
Y3+ (n=15)	39,022	02:01:57	2,601.47	0:08:08

When the data for the two groups was compared in terms of reading patterns, both groups engaged more in selective reading patterns than careful reading when reading the source texts (see Table 7). In terms of percentage, the Y3+ group appeared to have proportionally more selective reading patterns than the Y1 group. These differences perhaps suggest a greater reliance on selective reading skills on the part of more experienced students but this observation will need to be investigated in detail in future studies.

In summary, for RQ2, the data suggests that the reading-into-writing test task elicited all four reading patterns (careful local, careful global, selective local and selective global) from the participants. A key finding is that the task consistently elicited a much higher percentage of selective reading patterns than for careful reading: approximately 70% selective reading and 30% careful reading. As mentioned before, due to limited scope, we report only the eye-tracking data of the validation study in this chapter but it is useful to note that through triangulation of eye-tracking and interview data, it is possible to further differentiate types of selective reading. During the structured interview sessions after the task, participants reported using selective reading to skim for gist (skimming) and to scan for specific facts and information

Table 7 Comparison of reading types for Y1 and Y3+ participants

Y1 (n=15)					
Reading patterns	**Total no. of fixations**	**Total duration (hh:mm:ss)**	**Mean total no. of fixations per participant**	**Mean total duration per participant (hh:mm:ss)**	**Percentage of fixations**
Careful local	4,172	00:13:26	278.13	0:00:54	10%
Careful global	8,500	00:28:00	566.67	0:01:52	21%
Selective local	14,065	00:40:23	937.67	0:02:42	35%
Selective global	13,625	00:39:48	908.33	0:02:39	34%
Total	40,362	02:01:37	2,690.80	0:08:07	100%
Y3+ (n=15)					
Careful local	3,287	00:10:46	219.13	0:00:43	8%
Careful global	7,015	00:23:58	467.67	0:01:36	18%
Selective local	14,356	00:43:03	957.07	0:02:52	37%
Selective global	14,364	00:44:10	957.60	0:02:57	37%
Total	39,022	02:01:57	2,601.47	0:08:08	100%

(scanning). Participants also reported repeatedly re-reading certain small parts of text to facilitate note taking as well as improving comprehension (for a full account of the interview data, see Latimer 2018).

The findings imply that reading activities required on the reading-into-writing task differ considerably from the reading of isolated sentences studied in many previous studies (e.g. Ashby, Rayner and Clifton 2005, Rayner, Li, Williams, Cave and Well 2007). The progressive careful reading reported in much of the eye-tracking literature to date appeared to be only part of the academic reading construct. The results suggest that selective reading was used extensively for purposes such as reading for gist, search reading, scanning for information and intensive re-reading. One of Weir's concerns about students' reading skills was that most EAP and university admissions tests do not have a specific focus on selective reading skills (Weir, Vidaković and Galaczi 2013). For many students, the bulk of their assessments would be in the form of reading-into-writing tasks (Bridgeman and Carlson 1983, Hale et al 1995, Rosenfeld, Leung and Oltman 2001). Therefore, it would seem particularly important to be able to assess students in terms of their selective and careful reading skills. The development and application of tests of selective reading, in conjunction with tests of careful reading, would result in positive washback with students developing an awareness of the different types of reading and when and how they are utilised before they embark on their academic careers.

Personal reflections

Cognitive validity lies at the heart of Weir's work and therefore understanding how students utilised different types of reading when undertaking the BART test papers was a key concern for Cyril and his colleagues at CRELLA. Cyril was excited about the prospect of using technology to investigate students' use of different types of reading under real-life and test conditions. The *a priori* validation study reported was part of Nicola's PhD research. As a team (Cyril and Sathena were Nicola's supervisors), we had numerous discussions from conceptualising the design of the study to analysing eye-tracking data in relation to theories of reading. These findings, that selective or expeditious forms of reading play a major, perhaps even a majority, role when students complete a reading-into-writing task, lend further support to Cyril's life-long research on different types of reading.

Conclusion

There were several limitations to this validation study. Firstly, the task used in this study represented a very short reading-into-writing task and therefore may not have elicited the full range of processes demanded of students completing genuine coursework tasks. Secondly, the texts for this task were preselected, therefore participants did not have to evaluate the texts in the way they would normally do. Finally, although the algorithm used to categorise the data was developed carefully on the basis of both an understanding of the literature and interpretation of empirical data, there might be other ways of analysing the patterns of the eye fixations. Considering the limitations, the method proposed allows researchers to differentiate between students' eye-tracking evidence of careful reading and selective reading on a test. It is our hope that this would be useful for researchers who wish to interpret use of different types of reading from patterns of eye movement.

The results of the validation test provided insight into the reading patterns elicited by a reading-into-writing test task. It lends support to Weir's long-standing argument that readers make decisions about what to read and how to read, utilising different types of reading (careful, expeditious, local, global) to access and integrate the information they need to meet their reading goals. By presenting how the construct of academic reading was operationalised in a test within a local context, this chapter aims to illustrate the impact of Weir's work on shaping our understanding of the nature of academic reading (or more widely academic literacy skills) and the value of Weir's model of reading in providing a theoretical basis for test design and validation.

Endnote

Few people would disagree that Cyril was very knowledgeable. He often said that a researcher needs to 'know their stuff' and one should always 'read before they do anything'. At the time, little did we know that what he really meant was that we needed a thorough understanding of the past and present of language testing in the world. He insisted on the importance of having a historical and global perspective of language testing. Whilst he was eminently knowledgeable, he was quite forgiving of his students' ignorance. He encouraged and guided his colleagues and students to read widely and critically within and beyond our areas of interest. He used to keep his massive personal collection of books at CRELLA. He was always quick to recommend which books (sometimes the exact chapters or pages) we needed to read. Cyril supervised about 20 PhD students over the course of his career. He supervised both of us on our PhD journey and was a supportive, encouraging and inspiring mentor. He challenged his junior colleagues to question and criticise, and advised them to never accept any theory, however well established, unquestioningly. Cyril was particularly enthusiastic in encouraging his academic colleagues, of all levels, to criticise his own work. He prized academic rigour and that, perhaps, is one of the many reasons he was so esteemed in his field. He was keen to involve PhD students in real language testing projects and he trusted them with important responsibilities. He recognised people's potential (usually before they themselves did) and he was generous in helping them to unlock their potential. We were privileged to have known him. A great mentor and a dear friend we will miss.

References

Ashby J, Rayner, K and Clifton, C (2005) Eye movements of highly skilled and average readers: Differential effects of frequency and predictability, *Quarterly Journal of Experimental Psychology* 58A, 1,065–1,086.

Bax, S (2013) The cognitive processing of candidates during reading tests: Evidence from eye-tracking, *Language Testing* 30 (4), 441–465.

Bax, S and Chan, S (2019) Using eye-tracking research to investigate language test validity and design, *System* 83, 64–78.

Beard, R (1972) *Teaching and Learning in Higher Education*, Harmondsworth: Penguin.

Bråten, I and Strømsø, H (2003) A longitudinal think-aloud study of spontaneous strategic processing during the reading of multiple expository texts, *Reading and Writing* 16 (3), 195–218.

Bridgeman, B and Carlson, S (1983) Survey of academic writing tasks required of graduate and undergraduate foreign students, *ETS Research Report Series* 1983 (1), available online: onlinelibrary.wiley.com/doi/abs/10.1002/j.2330-8516.1983.tb00018.x

Brunfaut, T and McCray, G (2015) *Looking into test-takers' cognitive processes whilst completing reading tasks: a mixed-method eye-tracking and stimulated recall study*, British Council ARAGs Research Reports Online, available

online: www.britishcouncil.org/sites/default/files/brunfaut_and_mccray_report_final_0.pdf

Chan, S (2018a) *Defining Integrated Reading-into-Writing Constructs: Evidence at the B2–C1 Interface*, English Profile Studies volume 8, Cambridge: UCLES/Cambridge University Press.

Chan, S (2018b) Some evidence of the development of L2 reading-into-writing skills at three levels, *Language, Education and Assessment* 1, 9–27.

Davies, A (1984) Validating three tests of English language proficiency, *Language Testing* 1 (1), 50–69.

Ehrlich, S F and Rayner, K (1981) Context effects on word perception and eye movements in reading, *Journal of Verbal Learning & Verbal Behavior* 20, 641–655.

Fisher, D (1983) *An experimental study of eye movements during reading*, unpublished manuscript.

Hale, G, Taylor, C, Bridgeman, B, Carson, J, Kroll, B and Kantor, R (1995) A study of writing tasks assigned in academic degree programs, *ETS Research Report Series* 1995 (2), available online: onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.1995.tb01678.x

Hoover, W A and Tunmer, W E (1993) The components of reading, in Thompson, G B, Tunmer, W E and Nicholson, T (Eds) *Language and Education Library 4: Reading Acquisition Processes,* Clevedon: Multilingual Matters, 1–19.

Just, M A and Carpenter, P A (1980) A theory of reading: From eye fixations to comprehension, *Psychological Review* 87 (4), 329–354.

Khalifa, H and Weir, C J (2009) *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing volume 29, Cambridge: UCLES/Cambridge University Press.

Knoch, U and Sitajalabhorn, W (2013) A closer look at integrated writing tasks: Towards a more focussed definition for assessment purposes, *Assessing Writing* 18 (4), 300–308.

Latimer, N (2018) *Reading during an academic reading-into-writing task: an eye-tracking study*, unpublished PhD thesis, University of Bedfordshire.

Rayner, K (1977) Visual attention in reading: Eye movements reflect cognitive processes, *Memory & Cognition* 4, 443–451.

Rayner, K, Pollatsek, A, Ashby, J and Clifton, C (2012) *Psychology of Reading*, New York: Psychology Press (Second edition).

Rayner, K, Li, X, Williams, C C, Cave, K R and Well, A D (2007) Eye movements during information processing tasks: Individual differences and cultural effects, *Vision Research* 47 (21), 2,714–2,726.

Rosenfeld, M, Leung, S and Oltman, P K (2001) *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels*, TOEFL® Monograph Series MS-21, available online: www.ets.org/Media/Research/pdf/RM-01-03.pdf

Stanovich, K E (1980) Towards an interactive compensatory model of individual differences in the development of reading fluency, *Reading Research Quarterly* 16 (1), 32–71.

Starr, M and Rayner, K (2001) Eye movements during reading: some current controversies, *Trends in Cognitive Science* 5 (4), 156–162.

Tobii, A B (2018) *Tobii Pro X2-60 eye tracker*, available online: www.tobiipro.com/product-listing/tobii-pro-x2-30/

Urquhart, A and Weir, C J (1998) *Reading in a Second Language: Process, Product and Practice*, Harlow: Pearson Education Ltd.

Weir, C J (1983) *Identifying the language problems of the overseas students in tertiary education in the United Kingdom*, unpublished PhD thesis, University of London.

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

Weir, C J and Chan, S (2019) *Research and Practice in Assessing Academic Reading: The Case of IELTS*, Studies in Language Testing volume 51, Cambridge: UCLES/Cambridge University Press.

Weir, C J and Porter, D (1996) The multi-divisible or unitary nature of reading: the language tester between Scylla and Charybdis, *Reading in a Foreign Language* 10 (2), 1–19.

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

Weir, C J, Yang, H and Jin, Y (2000) *An Empirical Investigation of the Componentiality of L2 Reading in English for Academic Purposes,* Studies in Language Testing volume 12, Cambridge: UCLES/Cambridge University Press.

Weir, C J, Hawkey, H, Green, A, Ünaldi, A and Devi, S (2012) The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university, in Taylor, L and Weir, C J (Eds) *IELTS Collected Papers 2*: *Research in Reading and Listening Assessment*, Studies in Language Testing volume 34, Cambridge: UCLES/Cambridge University Press, 37–119.

Wu, R Y F (2014) *Validating Second Language Reading Examinations: Establishing the validity of the GEPT through alignment with the Common European Framework of Reference*, Studies in Language Testing volume 41, Cambridge: UCLES/Cambridge University Press.

Yu, G (2005) *Towards a model of using summarization tasks as a measure of reading comprehension*, unpublished PhD thesis, University of Bristol.

Zola, D (1984) Redundancy and word perception during reading, *Perception and Psychophysics* 36 (3), 277–284.

Appendix 1

Key stages of developing BART

- Reviewing the landscape: an extensive review of current academic reading tests.
- Weir's (1983) investigation into the language activities and associated problems of students studying at tertiary level throughout the UK, amongst others, fed into the development of the test specifications.
- Using the metrics established by Weir et al (2012), the texts used in the test papers were developed to reflect the level of difficulty of undergraduate reading texts.
- The tasks were designed to elicit the types of reading identified in Weir's reading model (Khalifa and Weir 2009) and Chan's reading-into-writing model (2018a, 2018b).
- Four versions of the test were developed.
- The rating scales were developed to reflect a componential approach, with separate scores for different types of reading on the reading paper and separate scores for content, organisation and language for the reading-into-writing paper.
- Feedback from the wider team at CRELLA was sought on the draft papers.
- A mini-pilot was conducted with 30 individuals from a range of courses. Analysis of feedback from participants suggested that the tests functioned well in terms of the overall layout of the various sections, timing and clarity of instructions.
- In Pilot 1, 218 year-one undergraduates completed 186 reading papers (evenly spread across four versions) and 217 reading-into-writing papers (version 1). Students were studying a range of courses including life sciences, computing, business, sports and nutrition, and child development.
- The test was adapted to enable the papers to be delivered via computer.
- In Pilot 2, 572 year-one students completed 519 reading papers (evenly spread across four versions) and 309 reading-into-writing papers (evenly spread across two versions). Students were studying a range of courses

including life sciences, computing, business, sports and nutrition, child development, education, and health and social care.

- The results from Pilot 2 are currently being subjected to item analysis to improve reliability and consistency.

6 Applying the socio-cognitive framework: gathering validity evidence during the development of a speaking test

Fumiyo Nakatsuhara

Centre for Research in English Language Learning and Assessment, Bedfordshire

Jamie Dunlea

Assessment Research Group, British Council

This chapter describes how Weir's (2005; further elaborated in Taylor (Ed) 2011) socio-cognitive framework for validating speaking tests guided two *a priori* validation studies of the speaking component of the Test of English for Academic Purposes (TEAP)[1] in Japan. In this chapter, we particularly reflect upon the academic achievements of Professor Cyril J Weir, in terms of:

- the effectiveness and value of the socio-cognitive framework underpinning the development of the TEAP Speaking Test while gathering empirical evidence of the construct underlying a speaking test for the target context
- his contribution to developing early career researchers and extending language testing expertise in the TEAP development team.

Introduction

The Test of English for Academic Purposes (TEAP) is a new admissions test for Japanese colleges and universities, whose administration with all four skills papers (i.e. Reading, Listening, Writing and Speaking) commenced in 2014. It is designed to measure the English language proficiency of Japanese upper-secondary school students intending to study at Japanese colleges and universities. Since its full launch in 2014, the number of test takers has increased rapidly year by year, and is predicted to be more than 30,000 in the 2019/20 academic year (Eiken Foundation no date). In the 2017/18 academic

1 This is distinct from the test designed originally by Cyril Weir in the 1980s named TEAP (Test *in* English for Academic Purposes) which was then was changed to Test in English for Educational Purposes (TEEP, still used by the University of Reading).

year, TEAP was delivered in 20 locations nationally and recognised by 120 universities (Eiken Foundation 2018). TEAP was from its outset intended to prove an innovative example of how to answer calls for the reform of Japan's English language education system (Dunlea, Fouts, Joyce and Nakamura 2019, Green 2014). With the promotion of four-skills tests in university entrance exams at the centre of government reform policy initiatives, it has the potential to make an important contribution to this debate.

TEAP was a collaborative test development project that involved three partners: 1) Eiken Foundation of Japan, the largest English examination board in Japan which administers the EIKEN English proficiency tests to over 2 million test takers a year, 2) Sophia University, one of the leading private universities in Japan, and 3) the Centre for Research in English Language Learning and Assessment (CRELLA) at the University of Bedfordshire in the UK, which provided specialist assistance to the project. The initial development of the receptive skills components was undertaken by the Japan-based partners, and the reading and listening test specifications were later refined and formalised by Taylor (2014) from CRELLA. The Japan-based partners had appropriate expertise and long experience in test development, particularly through the EIKEN suite of tests. These tests, however, had developed through a long interaction with the local educational community, and thus reflected established approaches in the context of Japan. TEAP was intended to introduce new approaches, particularly in relation to productive skills, which would contribute to the reform of the entrance exam system (see Dunlea et al 2019 for a discussion of design decisions for TEAP that were innovative for the local context). This was an important factor in the decision to look outward to external, international expertise, and the TEAP Writing Test development project was initiated in 2009, led by Professor Cyril J Weir from CRELLA (Weir 2014). The TEAP Speaking Test development project built on this collaboration, and began in 2010, going through several research phases before its first administration in 2014.

This chapter draws heavily on Nakatsuhara's (2014) project report which detailed the first two *a priori* validation studies of the TEAP Speaking Test and were guided by Weir's (2005) socio-cognitive validation framework. Given the long-term aim of TEAP to foster a positive impact on English education in Japan, including on the processes and systems for high-stakes speaking test development (Green 2014, Dunlea et al 2019), the socio-cognitive framework offered a very useful model of test development while gathering empirical evidence of the construct underlying a speaking test for the target context. In this chapter we have chosen to describe in some detail the validation research that underpinned the TEAP Speaking Test for several reasons. First, we believe it testifies to the effectiveness and practical value of the socio-cognitive framework approach to test development and validation which Cyril was so instrumental in developing and promoting over many years. Secondly, it demonstrates the practical value of such an accessible and transparent approach

when working in an international partnership within what may be a complex socio-political, educational and cultural context. Finally, we also reflect from a personal perspective upon Cyril's contribution to developing early career researchers and extending language testing expertise in the TEAP development team. The first two points – the contribution of the socio-cognitive framework to the development of TEAP, and its particular usefulness knitting together both international and local perspectives – will be clarified throughout the discussion of the studies underpinning TEAP Speaking described below. Along with the research narrative, our personal reflections on Cyril's contribution will be offered at relevant stages of the project.

Reflections on Cyril's contribution

First of all, it is worth noting the critical role that Cyril played in facilitating the relationships and project implementation that made these studies possible. The testing of speaking, as described further below, posed the greatest challenge in the development of a four-skills test in the context of Japan, and the Japan-based partners had been discussing how best to go about the development of the TEAP Speaking Test. Working with international partners for the writing test was a major change in approach, and would not have extended to the speaking component if the collaboration with CRELLA on TEAP Writing had not been successful. Cyril played a crucial role in establishing confidence and trust amongst the partners in the efficacy of inviting a foreign expert to lead on the academic design of a local speaking test. This extended beyond the substantive contribution of both the socio-cognitive model and Cyril's own extensive experience in test development. His personal style and leadership in building an atmosphere of mutual respect and confidence was invaluable.

The Japan-based partners were now open to continuing the international collaboration to address TEAP Speaking, and it was at this time that Cyril introduced to Eiken the first author of this chapter, at that time his junior colleague. Nakatsuhara had just completed her PhD on the use of speaking assessment with Japanese upper-secondary school students. Cyril mediated an initial meeting at the 32nd Language Testing Research Colloquium (LTRC) in Cambridge with the second author, at that time the Chief Researcher for Eiken. Cyril had the foresight to see that Nakatsuhara would not only bring the international expertise that CRELLA was being asked to provide, but she would also bring understanding of the local context and the motivation to contribute to positive change in a university entrance exam system through which she herself had passed. Cyril's confidence in Nakatsuhara's ability to lead on the speaking test development was integral to gaining the support of the other partners to allocate this important role to a new researcher, building and working within the framework of collaboration already established with Cyril for the TEAP Writing development.

As will be elaborated in the remaining sections of this chapter, the TEAP

Speaking project went through several research phases before its first administration in 2014, and it turned out to be a very fruitful and successful collaboration, from which all partners learned from each other, worked hard for the shared goal of developing such a speaking test that could contribute to improving the English education system in Japan, while making every possible effort to strike the most optimal balance between the best practice in speaking assessment and various practical constraints in the local context.

Background to the studies: Designing the TEAP Speaking Test

Sasaki's (2008) summary of the 150-year history of English language education and assessments in Japan highlights that greater emphasis is now placed on the teaching of speaking skills as *practical communication abilities*. The current course of study for upper-secondary schools (MEXT 2009) encourages the use of communicative speaking activities in the classroom. Innovations also include the obligatory status of a speaking component as a part of university admissions tests (Dunlea et al 2019). Nevertheless, despite these recent innovations, practical information on how to assess students' speaking abilities was not made sufficiently accessible to classroom teachers. Furthermore, it was left to local and international examination boards to develop and propose their speaking tests to be approved for use as university entrance tests. As such, a significant gap remained between policy goals and changes to actual practice on the ground. The TEAP project, therefore, had from the outset placed importance on creating positive washback (see Green 2014 and Dunlea et al 2019 for a comprehensive overview of the impact intended for TEAP), and the TEAP development team strongly hoped that the introduction of a standardised TEAP Speaking Test with transparent test specifications could help to promote the testing of speaking abilities in Japan, and to provide a transparent model for designing a speaking test suitable for the local context.

To achieve the long-term goal, a number of sources were carefully considered to inform draft test specifications of the TEAP Speaking Test, which we will briefly describe below.

Ministry of Education, Culture, Sports, Science and Technology (MEXT) guidelines

An initial background review was conducted by the Eiken project team. It examined the new curriculum guidelines for upper-secondary schools (MEXT 2009), regarding the types of compulsory and optional English modules, and language use situations and language functions to be focused on in these modules. This review provided valuable information for understanding trends in the Japanese education sector relevant to TEAP.

Socio-cognitive framework and literature review

From the outset of the project, Weir's (2005) socio-cognitive framework was selected to be used as a test development and validation framework for TEAP due to its enhanced practicality for test developers and its great transparency for test users. O'Sullivan and Weir (2011:20) describe the framework as 'the first systematic attempt to incorporate the social, cognitive and evaluative (scoring) dimensions of language use into test development and validation'. The framework consists of five validity components: *cognitive, context, scoring, consequential* and *criterion-related* validity, and it represents a unified approach to gathering validation evidence for developing and validating tests. Weir (2005) provides initial versions of the framework adapted for each of the four skills, and the framework for speaking has been applied and refined in Taylor (Ed) (2011). It is particularly valuable that the framework highlights the significance of providing evidence for *cognitive* and *context* validity during the initial test development stage (Taylor 2011:25–28). Dunlea (2015) furthers this concept of *a priori* validation by noting that all of the validity evidence categories are likely to be called upon in an integrated, iterative process of development, with some receiving more or less prominence depending on the purpose of data collection at each stage of development. Following Dunlea's (2015) reconceptualisation, all validity components played a role in the development of TEAP Speaking, with the main focus being cognitive, context and scoring validity.

As part of the first preparatory work undertaken prior to drafting test specifications and deciding on speaking test formats, Nakatsuhara (2010) provided a review of the assessment literature on speaking ability and of available speaking practices using the socio-cognitive validation framework. The review touched upon different aspects of validity while referring to how they relate to the target Japanese context and what critical questions the TEAP development team should be addressing in applying this framework to the development of the TEAP Speaking Test. The framework was instrumental in systematically organising the latest speaking assessment theories and research, comprehensively presenting various practices which reflect different speaking constructs, and synthesising the body of speaking assessment research and practice in an accessible and useful manner to the project team.

Common European Framework of Reference for Languages (CEFR)

The initial review also incorporated relevant CEFR descriptors (Council of Europe 2001) wherever appropriate. The CEFR played a central role in the whole TEAP project as a source for identifying criterial features of the different ability levels to be targeted by different test tasks. The CEFR descriptors

were also useful starting points for developing the necessarily more specific descriptors needed for use in rating scales. It was felt that bringing the CEFR into the test design from the beginning would facilitate stakeholders' understanding of the test scores and task requirements. It should also be useful to report scores not only as scale scores but in bands which can indicate to test takers their approximate level in terms of some external criterion, and the CEFR offered possibilities here. While the CEFR had been gaining traction for some time in Japan and other Asian countries as a useful framework to help set educational attainment goals and inform assessment reform, there were of course arguments for and against introducing an external 'international' framework into the local context of Japan. It is beyond the scope of this chapter to give a detailed overview of this debate, but it is worth noting that the decision to use the CEFR from the outset in the design of TEAP was informed by emerging empirical and conceptual research publications in the local context (e.g. Dunlea 2015, Dunlea and Figueras 2012, Dunlea et al 2019).

Following the decision made for TEAP Writing, it was decided that the TEAP Speaking Test should also be able to provide useful feedback to students at the A2 level of proficiency, as this is one of the benchmark levels of ability recommended by MEXT, and one that is probably closer to reality for a large number of upper-secondary school students. In this way, the TEAP programme from the outset placed the typical test takers at the centre of the test design, both in terms of what can realistically be expected of upper-secondary school students and providing useful feedback. At the same time, in order to look forward to the more demanding target language use (TLU) domain of the academic learning and teaching context of Japanese universities, it was decided that the test should contain tasks capable of discriminating between students at B1 level and the more advanced B2 level appropriate to the TEAP TLU domain, and be able to provide useful feedback for students at this more advanced level of ability.

Needs analysis

Given the role of TEAP as a university entrance examination and the importance of the test to reflect the TLU domain, it was considered critical for the test specifications to be informed by the language functions that upper-secondary school teachers wish their students to master by the end of upper-secondary education and that university teachers consider to be significant for a student to be successful in first-year undergraduate classes.

The former was carried out via a questionnaire survey with 172 upper-secondary school teachers using O'Sullivan, Weir and Saville's (2002) function checklist (Nakatsuhara 2010). The latter was conducted through a questionnaire survey with 24 English teachers at Sophia University who were teaching

first-year students at the time of the data collection. Both surveys utilised a comparable function checklist based on O'Sullivan et al (2002) with a list of over 20 functions sub-divided into informational functions (e.g. providing personal information, expressing opinions, justifying opinions), interactional functions (e.g. agreeing, asking for information, negotiating meaning) and managing interaction functions (e.g. initiating interaction, changing topics). Findings from both surveys are summarised in Nakatsuhara (2014).

Needs analysis had long been advocated as an essential component for test development (Weir 1983; see also Chapter 1, this volume). While the project timeline did not allow larger-scale needs analysis surveys, the findings offered valuable sources on which to base the selection of task types, in the attempt to make the test appropriate to the local context.

Iterative discussions among the project partners

Based on the above sources, an early version of the test specifications structured according to the socio-cognitive framework was drafted, while close communication was maintained among the project team via email and video-conferencing. A one-day face-to-face meeting was then held in March 2011, which included the key project staff members from all three project partners to fully agree on each point of the specifications. The points extensively discussed included:

- TLU domain
- ability levels targeted
- rating criteria and score reporting
- interlocutors' and raters' roles and training
- preparation of the test handbook
- test structure and timing
- sample tasks
- contextual factors needing special attention for each task
- cognitive demands needing special attention for each task
- the role of the CEFR and its relevant scales and descriptors.

When discussing the types of task, consideration was given to the cognitive demands that each task would make on test takers. Following Field's (2011) model of grading cognitive demands of speaking tasks, the development team paid attention to cognitive demands in relation to *conceptualisation* and *grammatical encoding*.

The degree of cognitive demand for conceptualisation was manipulated under two parameters: *provision of ideas* and *integrating utterances into a discourse framework*, while that for grammatical encoding was specified in the form of language functions to be performed by test takers (Field 2011). The results of the language function surveys were used in conjunction with this discussion to make an informed decision regarding task formats. Although some

project members were initially keen to include paired or group oral formats to elicit richer *interactional* and *managing interaction* functions, it was agreed that a role-play task, where test takers ask questions to and maintain communication with the examiner, would be more appropriate. The decision reflected the survey findings that highlighted the importance of the ability to ask for information and opinions. This was considered an innovative feature of the test with potential for fostering positive washback by giving test takers the role of leading the interaction rather than just responding to the examiner.

By the end of the one-day discussion, the development team had agreed on a draft test structure, illustrated in Table 1.

Table 1 Test structure*

Part	Task (Level)	Time	Cognitive demands: Grammatical encoding	Cognitive demands: Conceptualisation	Example topics
1	Interview (A2–lower B1)	2 mins	Providing specific personal information at different temporal frames (present, past, future)	*a) ideas* Low *b) discourse framework (I–C)* Low	Study, languages, career, upper-secondary school life, university life
2	Role-play (B1)	2 mins	Initiating interaction Asking for information/ opinions Commenting	*a) ideas* Low *b) discourse framework (C–I)* High	Interviewing an upper-secondary school teacher, interviewing a university student who has been back from study abroad
3	Monologue (B1–B2)	2 mins (inc. 30 secs for prep)	Agreeing/ disagreeing Justifying opinions Elaborating	*a) ideas* Mid *b) discourse framework (C)* Mid–High	A topic related to the one discussed in Part 2
4	Extended interview (B2)	4 mins	Expressing opinions Justifying opinions Comparing Speculating Elaborating	*a) ideas* High *b) discourse framework (I–C)* High	Two subject areas that are more topical and abstract than those in the previous parts, e.g. means of transportation, festivals, health, studying and travelling abroad, education system

**'Discourse framework' in column 5 refers to the way in which interaction is organised. '(I–C)' and '(C–I)' indicate interlocutor–candidate interactions led by the interlocutor and by the candidate, respectively. '(C)' indicates a monologic talk by the candidate. See Field (2011) for more information.*

Reflections on Cyril's contribution

Building upon the successful collaboration model of TEAP Writing that Cyril had already established with the Japan-based partners, the distance and face-to-face communications for TEAP Speaking went very smoothly and productively. While the three partners played complementary roles in the project and CRELLA was in charge of taking the academic lead in the TEAP productive skills projects, Cyril was always keen to listen to the local partners' needs and ideas, he never trivialised others' opinions, and he was exceptionally good at making all project members feel valued in their contributions. The TEAP Speaking project built on this spirit of open discussion, enabling all project members to feel a sense of ownership of the project. Everyone's ideas were treated as equally important and fully discussed in light of the shared goal of developing a speaking test which has a sound theoretical foundation, and which would foster positive washback in the English educational system in Japan. The project team also carefully examined the logically complex nature of speaking assessment and how best available resources can be utilised, since the socio-cognitive framework placed great importance on bridging theories and practice, reflecting Cyril's belief that a validation framework has to be useful to the practitioners who have to work with all sorts of practical constraints and limited resources.

Focus group discussions and a mini trial

Once the task types and the rating categories were agreed, it was time to draft the rating scales. Given the vital role of the CEFR in designing the TEAP test, the CEFR descriptors from the most relevant scales were used as the criterion benchmarks from which TLU-specific descriptors for TEAP Speaking were developed. This was done with the explicit intention of building the CEFR into the rating scales and test design for the purposes of reporting the results to test takers. In addition, other established rating scales such as the Cambridge ESOL Common Scale for Speaking (Galaczi, ffrench, Hubbard and Green 2011), and those developed for Japanese learners of English for the Standard Speaking Test (SST) (ALC 2006) and Kanda English Proficiency Test (KEPT) (Bonk and Ockey 2003) also informed the rating scales development.

An early draft version of the rating scales was then discussed in a focus group within each of the project partners individually followed by a larger focus group among all three partners. These discussions were also informed by video recordings gathered in a mini-trial test with three first-year university students who were at approximately A2, B1 and B2 levels. The discussions were repeated several times until they reached an agreed version. The draft scales contained five analytical categories (*Grammatical range and accuracy*, *Lexical range and accuracy*, *Fluency, Pronunciation*, and *Interactional*

effectiveness), each of which had four levels (0=below A2, 1=A2, 2=B1 and 3=B2).

As described thus far, the development of draft specifications for the TEAP Speaking Test was informed by various sources. With the draft specifications, two *a priori* validation studies were carried out in July 2011 (Study 1) and December 2011 (Study 2).

A priori validation studies

Weir (2005) asserted that establishing validity evidence should start at the before-the-test event stage, and the socio-cognitive framework was designed to guide how this can be achieved systematically. The studies presented here did so in two stages. Study 1 examined how well the draft test materials and rating scales operationalised the test construct in terms of certain aspects of *context* validity (which also gave some indication of the *cognitive* demands placed on the test takers) and *scoring* validity. Study 2 then investigated how well the test functioned in terms of *scoring* validity after incorporating the modifications suggested by Study 1. Due to space limitations, we will report only selected parts of the two studies (for a full range of research questions, methodology and findings, see Nakatsuhara 2014).

Research questions

Research questions (RQs) addressed through Studies 1 and 2 were:
RQ1: To what extent does the test elicit intended language functions in each task? (Study 1)
RQ2: Is there any evidence from test takers' output language that validates the descriptors used to define the levels on each rating scale? (Study 1)
RQ3: What are the participating interlocutors', raters' and students' perceptions of the testing procedures? (Study 1)
RQ4: How well does the test function in terms of scoring validity, after incorporating modifications suggested in Study 1? (Study 2)

Methodology

Participants

Study 1: The first study involved 23 university students, three trained interlocutors and three trained raters. The 23 students were recruited from different English classes at Sophia University, to cover a wide range of proficiency levels. They were first-year students, who had spent only three months at Sophia University at the time of the data collection. The three interlocutors were English teachers at Sophia University, who attended an interlocutor

training session prior to the test event. It was considered that the profiles of the three interlocutors would reflect those of prospective interlocutors in the operational TEAP Speaking Test. The three raters were selected by Eiken. All raters were experienced teachers at Japanese universities but with different levels of experience as raters in standardised speaking tests. They all attended a rater training session prior to the test.

Study 2: A total of 120 third-year upper-secondary school students were recruited to take part in the second study. Five interlocutors were involved, of whom four were English teachers at Sofia University, and one was a trained and experienced rater for the EIKEN Speaking tests. Six raters participated in Study 2, who were all native speakers of English. They were fairly experienced teachers at Japanese universities and/or lower-secondary and upper-secondary schools, as well as being experienced raters in standardised speaking tests.

Data collection

Study 1: During the Study 1 data collection, the three trained interlocutors interviewed 23 test takers, but they did not assign marks to test takers' live performances. After completing their test sessions, they were asked to fill in a feedback questionnaire about different aspects of the interlocutor frame and interviewing procedures. Similarly, students also completed a feedback questionnaire about their test taking experience immediately after their participation. All performances were video-recorded, and all 23 recorded performances were rated by the three trained raters using the draft rating scales described earlier. Their feedback was also gathered by a post-rating questionnaire.

Study 2: The 120 students were interviewed by five interlocutors using modified test materials based on Study 1 findings. All speaking test sessions were video-recorded, and the six raters independently rated 60 video-recorded performances each. The rating followed a matrix to ensure sufficient overlap to enable the analysis of the data with multi-faceted Rasch analysis.

Data analysis

In preparation for the language function and micro-linguistic analysis described below, all Study 1 video recordings were transcribed using a simplified version of Conversation Analysis (CA) notation (Atkinson and Heritage 1984). CA transcription allowed for examining micro-analytic features of interaction between the examiner and the test taker.

Language function analysis (RQ1)

The transcripts were firstly analysed for the coverage of language functions elicited in each task. O'Sullivan et al's (2002) function checklist was slightly modified for use with the given data. While the checklist was originally developed for analysing language functions elicited in paired speaking tasks

of Cambridge English exams, the potential to apply the list to other speaking tests such as the IELTS Speaking Test (Brooks 2003) and the Graded Examinations in Spoken English (GESE) (Nakatsuhara 2018) has been explored. Since the list draws on Bygate's (1987) speaking model, the applicability of the checklist is not limited to any particular types of L2 test takers' speech, and was also useful to examine a range of language functions elicited in the TEAP Speaking Test.

This was the same list used for the language function surveys with educators at Japanese upper-secondary schools and Sophia University, which informed our selection of the task formats in the test. Therefore, the use of this checklist in this validation study enabled us to directly compare language functions specified in the test specifications reflecting the survey results with functions that are actually elicited from target test takers.

Linguistic and discourse analysis of students' speech samples (RQ2)

This analysis aimed at examining whether test takers' output language validates the descriptors used to define the levels on each rating scale. Previous studies have employed this approach to rating scale validation, including Brown (2006), and Iwashita, Brown, McNamara and O'Hagan (2008). A variety of linguistic measures were selected to reflect the features of performance relevant to the test construct defined within the draft analytical rating scales, so as to investigate whether these measures differ in relation to the proficiency levels of the test takers assessed using the rating scales. The transcripts were coded for these features by a research assistant. As with transcription, an interactive consensus approach to coding was taken. The project member who oversaw the data preparation reviewed several complete transcripts after they had been coded, and any differences in interpretation were resolved through discussion between the research assistant, the consultant and the project member overseeing the data preparation.

Three trained raters rated the 23 students' video-recorded test sessions, using the draft TEAP Speaking rating scales that consist of the following five categories: a) Grammatical range and accuracy, b) Lexical range and accuracy, c) Fluency, d) Pronunciation, and e) Interactional effectiveness. Since it is crucial that speech samples selected for the analysis are reliable representatives of a particular level of each analytical category, the test scores were first analysed using multi-faceted Rasch analysis.

Once the score analysis had confirmed the satisfactory and consistent ratings of the three raters, the video-recorded speech samples and their transcripts were analysed for the linguistic characteristics illustrated in Table 2. These linguistic features were selected to reflect elements of performance covered in the draft rating scale descriptors, except for the last three measures for the amount of talk. The linguistic features were analysed to investigate the extent to which each of these features differs between adjacent levels of the rating scales. Since not all measures were relevant for all parts of

the test, appropriate parts were selected for different analyses. For detailed explanation of how each of the characteristics was selected and measured, see Nakatsuhara (2014).

Table 2 Linguistic measures

Corresponding rating category	Focus	Measure	Parts of the test applied
a. Grammatical range and accuracy	Complexity	Ratio of subordinate clauses to AS-units	1, 2, 3, 4
		Number of words per AS-unit	1, 2, 3, 4
	Accuracy	Percentage of error-free AS-units	1, 2, 3, 4
b. Lexical range and accuracy	Range	Lexical frequency coverage (K1 + K2 + off-list words)	1, 2, 3, 4
		Academic Word List coverage	1, 2, 3, 4
	Accuracy*	—	—
c. Fluency	Hesitation	Number of unfilled pauses (utterance initial) per 50 words	1, 2, 3, 4
		Total pause time as a percentage of speaking time	3
	Disfluency	Ratio of repair, false starts, and repetition to AS-units	1, 2, 3, 4
	Temporal	Speech rate in Part 3	3
		Articulation rate in Part 3	3
d. Pronunciation	L1 influence	Number of words pronounced with noticeable L1 influence (katakana-like) as percentage of total words produced	1, 2, 3, 4
e. Interactional effectiveness	Length of response	Average words per response	1, 4
	Number of extra questions	Number of separate questions asked that were not on required list in Part 2	2
	Back-channelling and comments	Number of instances of back-channelling and comments in Part 2	2
f. Others – the amount of talk	Length of long turn	Total number of words produced in Part 3	3
	Total production	Total amount of production across all parts of the test, measured in words	1, 2, 3, 4
		Total number of AS-units produced across all parts of the test	1, 2, 3, 4

* *Lexical accuracy was not measured in this analysis, as it was deemed impossible to reliably identify word choice errors.*

Questionnaire analysis (RQ3)

Students', interlocutors' and raters' responses to feedback questionnaires were analysed using descriptive statistics. All question items were accompanied with comment boxes where the respondents could elaborate on their dichotomous or Likert-scale responses. Comments provided for each question were used to interpret and elaborate on the statistical findings. The feedback questionnaire given to students included questions regarding clarity of the task instructions, appropriacy of the speaking time duration, appropriacy of topics and task types, and comfort of physical testing conditions. Questions explored with the interlocutor questionnaire included appropriacy of time allocation, task prompts, main and follow-up questions, and easiness of time management and test administration procedures. The rater questionnaire explored questions such as clarity and usefulness of the rating scale, quality of the video recordings, quantity of ratable language elicited, and their rating processes.

Score analysis (RQ4)

Of the 120 students recruited, 113 students' scores were analysed due to seven absences on the day of testing. Multi-faceted Rasch analysis was carried out using the Facets program (Linacre 2011) for three major facets for the score variance in this study: examinees, raters and rating categories. The Partial Credit Model was used for the analysis.

Reflections on Cyril's contribution

The overall framework of Studies 1 and 2 of the TEAP Speaking project described so far had a deal of synergy with the TEAP Writing project, while specific research methods within each study were informed by relevant research literature to suit the construct that was targeted in TEAP Speaking or Writing respectively. Throughout the TEAP Writing and Speaking projects, a close dialogue was maintained across the two parallel projects. The spirit of open discussion that was noted earlier was pertinent not only across the project partners but also between the two external consultants. Throughout the TEAP Speaking project, Cyril generously guided Nakatsuhara, for whom this was the first large-scale international research consultancy, advising her on all aspects of project management, successful consultancy and international collaboration. With his support, the TEAP Speaking project progressed smoothly, and by the time Study 2 was designed, the progress of the TEAP Speaking project had caught up with that of the Writing project, enabling the project team to carry out Study 2 of the parallel projects with the same participants. Cyril prioritised soliciting ideas and information from the project partners as equal collaborators, and in addition created an atmosphere at CRELLA which valued the advice and suggestions of all, including junior colleagues. This project was no exception. Cyril approved of the way in

which Nakatsuhara had designed Study 2 of the TEAP Speaking project, and he decided to adopt a similar design in Study 2 of the Writing project.

Analysis and results

Language functions

Table 3 shows the target language functions in each task and average number of turns in which each function was produced per participant across the four parts of the test. Functions with an average realisation rate of 0.7 turns or above per test taker are in bold, based on the project team's agreement on the threshold for identifying the main functions elicited in the test (i.e. a function being elicited from 70% or more test takers).

Table 3 Language functions targeted and elicited across the four parts of the test

	Part 1		Part 2		Part 3		Part 4	
	Target	Mean	Target	Mean	Target	Mean	Target	Mean
Informational functions								
giving personal info (present)	✓	**1.70**		0.00		0.00		**1.04**
giving personal info (past)	✓	**2.30**		0.04		0.17		0.39
giving personal info (future)	✓	**1.74**		0.13		0.00		0.04
expressing opinions/ preferences		**4.52**		0.00		0.04	✓	**5.09**
elaborating		**2.52**		0.00	✓	**0.74**	✓	**2.17**
justifying opinions		0.65		0.00	✓	**1.74**	✓	**3.35**
comparing		0.00		0.00		0.00	✓	**2.30**
speculating		0.00		0.00		0.00		**0.87**
staging		0.00		0.22		0.04		0.00
describing a sequence of events		0.00		0.00		0.00		0.00
suggesting		0.00		0.00		0.00		0.04
Interactional functions								
agreeing		0.00		0.00	✓	**0.96**		0.13
disagreeing		0.00		0.00	✓	0.17		0.09
modifying		0.00		0.00		0.00		0.00
asking for opinions		0.00	✓	**1.87**		0.00		0.00
asking for info		0.22	✓	**3.87**		0.00		0.04
commenting		0.00	✓	**1.70**		0.00		0.09
asking for permission		0.04		**0.83**		0.00		0.00
greeting		**2.04**		**0.83**		0.00		0.39

Table 3 (continued)

	Part 1		Part 2		Part 3		Part 4	
	Target	Mean	Target	Mean	Target	Mean	Target	Mean
thanking		**0.78**		**1.30**		**1.09**		**2.48**
Negotiating meaning – check understanding		0.43		0.39		0.09		**0.74**
– indicate understanding		0.04		**1.00**		0.04		0.17
– ask for clarification		0.17		0.09		0.00		**0.74**
– correct others		0.00		0.00		0.00		0.00
– respond to a clarification request		0.00		0.17		0.04		0.04
Managing interaction functions								
initiating		0.00	✓	**1.00**		0.00		0.00
changing		0.00		0.04		0.00		0.00
reciprocating		0.22		**2.22**		0.04		0.00

The analysis demonstrated that there were clear differences between the four parts in their capability of eliciting different types of function.

As intended in the test specifications, Part 1 of the test (interview) mainly elicited informational functions such as *giving personal information* in different temporal frames, *expressing opinions/preferences, elaborating*, as well as some interactional functions like *greeting* and *thanking*.

In contrast, language functions elicited in Part 2 (role-play) were characterised more as interactional, such as *asking for opinions, asking for information, commenting, asking for permission, greeting, thanking,* and *negotiating meaning (indicating understanding)*. The elicitation of language functions to manage interaction, like *initiating interaction* and *reciprocating*, was also noticeable. An example excerpt of asking for information (line 2) and commenting (line 4) is shown in Excerpt 1.

Excerpt 1: asking for information and commenting in Part 2[2]

1 **S2-1:** Ah. (0.8) And uh (1.1) do you have (.) problem in the class?
2→ **E:** Yes, students get sleepy [in the afternoon. ((laughs))
3 **S2-1:** [Ah.
4→ **S2-1:** I- I always (0.4) sleep in the afternoon

Part 3 of the test (monologue) elicited a limited number of language functions. However, language to *agree/disagree* and to *justify opinions* expected from the task requirement was successfully observed.

2 **Transcription symbols:** (a) Unfilled pauses or gaps, periods of silence, and micro-pauses (less than 0.3 seconds) are shown as (.); longer pauses appear as a time within parentheses. (b) Dash - : cut-off. (c) Open bracket [: Beginning of overlapping utterances.

Part 4 of the test (extended interview) elicited a number of informational functions, such as *giving personal information (present), expressing opinions/preferences, elaborating, justifying opinions, comparing* and *speculating*. Test takers also *negotiated meaning (checking understanding/asking for clarification),* and *thanked* the interviewer, both of which are interactional functions.

The data confirms that the types of function observed in each part are congruent with the goals of each part, fully covering the functions described in the draft test specifications. It was also encouraging to find evidence that targeted language functions were not only elicited but were also elicited in ways that the test designers intended (for example transcripts, see Nakatsuhara 2014). This indicates that the intended constructs of the four tasks are appropriately operationalised. The analysis however suggested minor modifications to the interlocutor frame regarding rephrasing some questions in Part 1, limiting interviewers' response tokens in Part 3 to non-verbal ones, and standardising the wording to round off the Part 4 interaction.

Linguistic and discourse features

Due to space limitations, we will exemplify only selected measures for the validation of the draft 'Fluency' rating scale (for all analyses of the measures included in Table 2, see Nakatsuhara 2014).

Key assessment features specified in the draft Fluency scale were hesitation, disfluency features such as reformulation, and speed of speech. After reviewing the literature on measuring fluency (e.g. Inoue 2013, Iwashita et al 2008, Kormos and Dénes 2004, Tavakoli and Foster 2008, Wigglesworth and Elder 2010), it was decided to use two measures for hesitation, one measure for disfluency, and two measures for speed fluency. Of the five measures, we now present findings from one hesitation measure, one disfluency measure and one speed measure.

One of the measures for hesitation was the number of unfilled pauses per 50 words in all four parts. It was measured by the number of pauses of 0.3 seconds or longer which occurred after an examinee had begun speaking, divided by the number of words and multiplied by 50.

Disfluency was measured by the total number of features coded as instances of repair, false starts or repetition divided by the number of AS-units across the four parts. To do so, these disfluency features were firstly coded on the transcripts manually. Previous studies have used different formulations for disfluency analysis (e.g. Iwashita et al 2008), but the present study used the ratio to AS-units, as it was considered to represent more accurately the extent to which repair (dis)fluency would affect the message conveyed by the test takers.

For one of the speed measures, articulation rate was calculated, which was computed by the total number of syllables divided by the total duration of pure speech time. The measure was applied only to Part 3. This is because in

interactional parts of the test (Parts 1, 2 and 4), it is not possible or desirable to determine the ownership of unfilled pauses between turns; that is, both conversants (i.e. interviewer and test taker) are responsible for such pauses unless the previous speaker nominates the next speaker. Table 4 and Figures 1 to 3 show the results of the selected three measures.

Table 4 Selected fluency measures across the three proficiency levels

Focus	Measure	Parts applied	Level	N	Min	Max	Mean	SD
Hesitation	Number of unfilled pauses per 50 words	1, 2, 3, 4	Level 1 (A2)	7	13.95	32.32	22.44	5.71
			Level 2 (B1)	11	9.91	24.83	17.44	4.44
			Level 3 (B2)	5	6.69	18.02	10.89	4.23
Disfluency	Ratio of repair, false starts, and repetition to AS-units	1, 2, 3, 4	Level 1 (A2)	7	0.86	2.19	1.43	0.53
			Level 2 (B1)	11	0.61	2.22	1.27	0.52
			Level 3 (B2)	5	0.53	1.08	0.80	0.19
Speed	Articulation rate	3	Level 1 (A2)	7	1.18	3.13	2.32	0.62
			Level 2 (B1)	11	2.15	3.23	2.88	0.35
			Level 3 (B2)	5	2.63	3.32	3.02	0.27

Figure 1 Number of unfilled pauses/50 words (all parts)

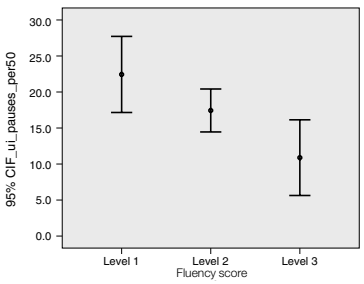


Figure 2 Ratio of repair, false starts, and repetition to AS-units (all parts)

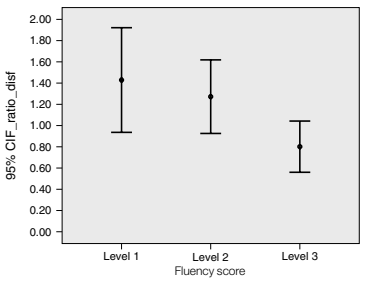
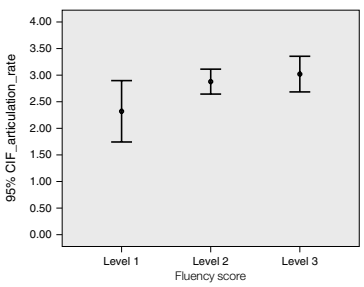


Figure 3 Articulation rate (Part 3)



The means of the three groups on all fluency measures varied in accordance with the rating scores that test takers obtained. In terms of hesitation, the number of unfilled pauses by Level 3 test takers was on average 10.89, while Level 2 test takers had 17.44 pauses and Level 1 test takers 22.44 pauses.

The ratio of disfluency features to AS-units clearly increased as the fluency scores decreased; Level 3 test takers showed on average one disfluency feature in four out of five AS-units (0.80), while Level 2 test takers had 1.27 features per AS-unit and Level 1 test takers had 1.43 features per AS-unit.

As for speed fluency, across the three proficiency levels, articulation rate in Part 3 changed in the expected direction. Level 3 test takers on average articulated 3.02 syllables per second, Level 2 test takers articulated 2.88 syllables, and Level 1 test takers articulated 2.32 syllables.

As such, we quantified linguistic and discourse features of test taker output that are related to key assessment features specified in the draft analytical rating scales: grammatical range and accuracy, lexical range and accuracy, fluency, pronunciation, and interactional effectiveness. In general, all examined features of test taker output varied according to the assessed proficiency level (Level 1, Level 2 and Level 3). All measures broadly exhibited changes in the expected direction across the three levels, providing the evidence that the rating scales are differentiating test takers' performance in a way congruent with the test designers' intention.

However, for a few measures, the difference between two adjacent levels was not as expected. For example, performances on the two adjacent levels were almost identical, or the differences between levels were greater at one boundary than the other. This result is in accordance with previous research (e.g. Brown 2006, Pollitt and Murray 1996), indicating that specific aspects of performance are probably more relevant to differentiate particular levels. This finding is worth following up to better understand the nature of test taker performance in the TEAP test. We should also bear in mind that it is necessary to replicate this study with a larger dataset, as the small sample size of the study did not allow for inferential statistics. Nevertheless, this study still offered useful *a priori* validity evidence early on to examine the extent to which the test developers' intentions were being operationalised, and to determine desirable modifications. Such information was intended to inform refinement of the specifications and scales before larger-scale piloting in Study 2.

Feedback

Selected responses to the student, interlocutor and rater questionnaires are reported below.

Student feedback

It is encouraging that 95.7% of the students found all test instructions clear and that 73.9% found all interviewer questions clear.

While the Part 2 role-play task is a new test task type in the Japanese context, 73.9% found it comfortable to attempt the Part 2 task, while some students wanted to have more freedom in thinking about their own questions ('I would have felt more comfortable if I had been asking my own questions only, rather than following a given list of questions'). 95.7% thought that the task reflected their real-life language use situations. Their comments included: 'In English classes, we often have to ask questions to native speaker teachers in English, and it is in fact very important to be able to ask questions in English'.

The length of preparation time in Part 3 was recognised as appropriate by 73.9% of the students, and 78.3% and 91.3% thought the topics were relevant for the third-year upper-secondary school students in Part 3 and Part 4, respectively.

The physical distance between the interviewer and the test taker was perceived as appropriate by 95.7%. It seems that students had split opinions about the beep sound of the timer – 60.9% thought that having the sound was good, but 39.1% disagreed with having the sound because 'the timer sound made me nervous, and instead the interviewer can just let the test taker know that the time is over'. Finally, it was very encouraging to find that none of the students thought that video-recording distracted their attention during the speaking test.

Interlocutor feedback

All three interlocutors in general felt the task timings, instructions, questions and general test administration were appropriate. There were some suggestions for improvement and comments on each question, such as:

- the question sequence should be more natural in the Part 1 interview
- the Part 2 role-play instructions should be clearer
- the use of a timer needs practising
- it is inevitable to deviate from the interlocutor frame in minor ways and the use of body language and back-channelling should be clarified.

Rater feedback

It seems that raters had some difficulties when they actually applied the scales to test taker performance. The 'Interactional effectiveness' category was perceived as the most difficult to use, receiving comments such as 'too long and confusing', and '[Fluency and Interactional effectiveness categories] are very hard to rate, as they are the result of a holistic impression'.

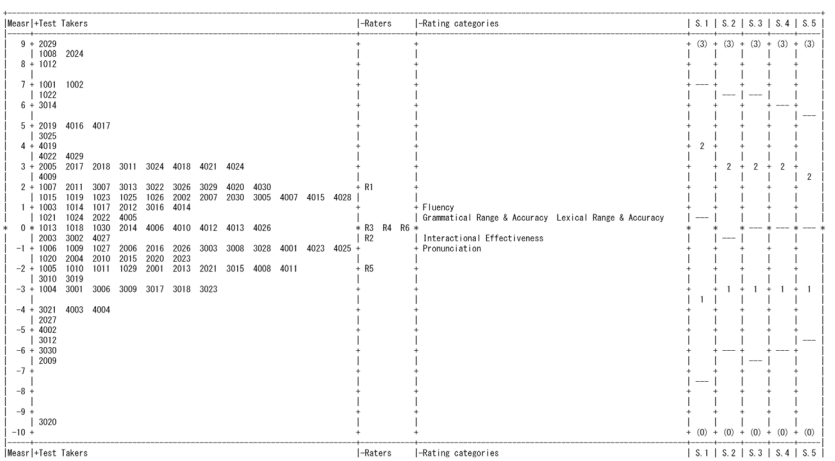
Two of the three raters needed to watch the video samples more than once to rate them but just to check part of the performance. It was generally encouraging that all three raters felt that the format provided a sufficient sample of language to distinguish between the intended levels. The rating process that the three raters followed varied. Two raters had similar processes: 'Hypothesis tested for each category during each part of the test. Made final decision at end'. By contrast, one had a fixed order in rating: 'Pronunciation, Lexis, Grammar, Fluency and finally Interactional effectiveness'.

As described so far, Study 1 investigated various aspects of the context and scoring validity of the TEAP Speaking Test to gather information on the extent to which the test materials and rating scales operationalised the test construct described in the draft test specifications. Based on the findings, several modifications were made to the test materials. They include: rephrasing one of the Part 1 questions, standardising interviewer behaviour in Parts 3 and 4, and adjusting the wording of descriptors in the Interactional effectiveness scale. Using the modified test materials, Study 2 was carried out to examine the scoring validity of the test.

Test scores

Figure 4 shows the overview of the results of the multi-faceted Rasch analysis, plotting estimates of examinee ability, examiner harshness and rating scale difficulty.

Figure 4 Overall Facets map (Study 2)



Examinees

The test discriminated well between examinees. The fixed (all same) Chi-square test was statistically significant ($\chi^2(112)=2{,}094.9$, $p<.005$). The separation index was 4.00, and the examinees were able to be separated into 5.67 statistically separate strata. The reliability was 0.94. The ability to separate the examinees into statistically distinct strata is important for the TEAP test since it is used for entrance purposes to discriminate between students of different ability levels.

For fit analysis, we followed Wright and Linacre's (1994) suggestion that infit mean square values in the range of 0.5 to 1.5 are 'productive for measurement'. Of the 113 students analysed, three students were identified as misfitting. The percentage of misfitting students in the dataset was 2.7%, almost satisfying McNamara's (1996:178) expectation that any test development should aim at having misfitting students at or below 2%.

Raters

Fit statistics on raters showed quite good fit of the six raters, indicating that they performed with a satisfactory degree of consistency. However, it was found that the six raters differed in terms of severity, and these differences were statistically significant ($\chi^2(5)=398.2$, $p<.005$). While the severity range was rather small for four out of the six raters, the difference between the harshest rater (i.e. Rater 1) and the most lenient rater (i.e. Rater 2) was 0.81 of a band. We suggested that these two raters should be retrained.

Rating categories

None of the rating criteria was misfitting. This was an encouraging result, as this indicates that the assumption of unidimensionality holds for this data (Bonk and Ockey 2003). This means that the separate analytic rating scales seem to be contributing to a common construct of 'speaking ability'. This is vital for the TEAP Speaking Test, which aims to provide a composite score by summing scores across the separate analytic scales.

While the analysis showed that the five rating categories exhibited significantly different degrees of difficulty ($\chi^2(4)=114.8$, $p<.005$), examination of probability curves for each category (see Figures 5 and 6 for two examples) demonstrated that the scale steps of all rating categories progressed in the order as designed, with each step being progressively more difficult than the lower step on the scale.

Figure 5 Grammatical range and accuracy scale

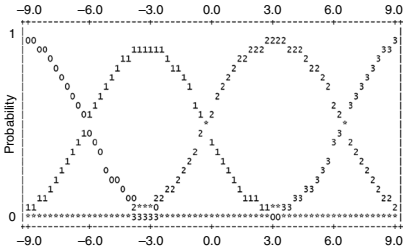
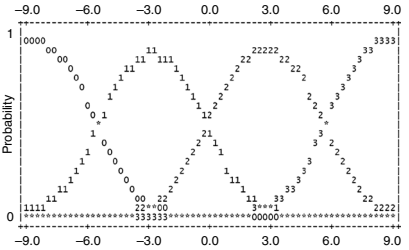


Figure 6 Interactional effectiveness scale



Discussion and conclusions

As already noted, it was beyond the scope of this chapter to provide a detailed description of all aspects of the TEAP Speaking Test development project (for a detailed description of the two *a priori* validation studies, see Nakatsuhara 2014). Our purpose was to highlight through an illustrative example how Weir's (2005) socio-cognitive framework provided accessible, theory-driven guidance for designing a speaking test appropriate for the high-stakes context of university entrance in Japan. As noted in Dunlea (2015), the evidence categories in the framework can be related to the six aspects of validity described in Messick (1996), and so provide the necessary evidence to 'touch all the bases' required to provide a robust justification for test use and interpretation that Messick ascribed to his model. Drawing on the framework, the developers were able to collect and integrate evidence in order both to give the test developers confidence they were travelling in the right direction, and to provide evidence to satisfy local stakeholders as well as international standards of best practice.

The iterative nature of the studies described highlights the importance of test developers being open to change and adaptation. The *a priori* validation studies described here did not just set out to confirm an initial design, but instead were intended to, and did, lead to refinement and improvements. The process underscored the importance and value of undertaking *a priori*, principled validation studies as a part of test development. At the same time, that iterative process also showed that the distinction made in the original 2005 framework of particular types of evidence into *a priori* or *a posteriori* phases did not hold in practice. All types of evidence were indeed relevant, but to different degrees in these *a priori* studies, as the developers built up a comprehensive picture of how the test worked in practice, and how that related to the original test design. The studies described here also highlighted the importance of giving contextual and cognitive aspects of the framework a central place at the very beginning of the design phase.

It is worth reiterating the closing comments of Nakatsuhara (2014:75) that 'the project offers a model for collecting different types of *a priori* validity evidence during the development stage of a speaking test'. Throughout the process, a balance was sought between practicality, local stakeholder needs and international best practice. It is not just the test that was offered as such a model, but the collaborative, iterative process of development and the theory that underpinned it.

Personal reflections

In addition to Cyril's contribution in providing the theoretically sound and practically useful validation framework which laid a solid foundation for the TEAP Speaking Test, we would like to conclude this chapter by reiterating personal reflections of him in relation to this project.

I joined CRELLA at the University of Bedfordshire led by Cyril (see Chapter 8, this volume) soon after I had completed a PhD. As noted earlier, this was the first large-scale research consultancy that I carried out in my career. Nevertheless, Cyril had confidence in me conducting this important project, and generously provided guidance throughout the course of the project. He was always available to offer his support whenever I had any academic and practical questions. He taught me a great deal about how best we can develop and validate tests, bridge theory and practice, manage a large-scale project and collaborate successfully with international partners. I learned that creating the atmosphere where every single project member can freely share their ideas and is willing to learn from each other is one of the most critical parameters of successful collaboration. I always admired him for being so humble to listen to his junior colleagues and other researchers and practitioners, despite him being such a well-known expert in the field. He was always there to help his colleagues, and he cared about his junior colleagues' careers more than they did for themselves. It is one of the many lessons he taught us that senior researchers should guide and help raise early career researchers for the future of the language testing field. (*Fumiyo Nakatsuhara*)

A defining memory of my experience of working with Cyril on this and other projects was the generosity of professional spirit he showed. He was committed to leaving behind not just a test or a validation study, but also the capability in the team members themselves to replicate, and indeed change, the methods he had introduced. Cyril approached the projects with an open mind and flexibility, and a willingness not only to adapt to the local needs which he was at pains to understand, but to learn from them and take away something new himself. This approach made complex theory accessible to those of us still developing our professional knowledge and skills. A colleague and key TEAP project member noted that theory sometimes comes with a bad reputation for those tasked with the hands-on job of making

something work in practice. The framework and the attitude and support that Cyril brought to the project put theory at the beginning, middle, and end of the process, but that theory worked and made a very difficult goal achievable and the results defensible. As noted above, the collaboration with CRELLA was a new approach for the Eiken team. This was an experiment not without risk for such an important and high-profile project. Cyril's international reputation encouraged the partners to make the decision to take that risk. His considerable experience, the strength of the socio-cognitive framework, and his approach to test development helped ensure the positive outcomes of the collaboration, and evaluation of it by the Japan-based partners. But reiterating the theme that we have touched on throughout our reflections, it was very much his personal values and respect for the ideas and well-being of all members of whatever project he engaged with that cemented the success of this international collaboration, which grew out of the Writing project and led to the studies described here. I suspect that Cyril would have been disappointed if he felt that he was *not* learning anything new from us, rather than just disseminating his already considerable expertise. But it is that, not always common, attitude which brought out the best from all of those of us lucky enough to have worked with him, and indeed in the studies we have described in this chapter. (*Jamie Dunlea*)

References

ALC (2006) *Standard speaking tests,* available online: www.alc.co.jp/edusys/sst/english.html

Atkinson, J M and Heritage, J (1984) *Structures of Social Action,* Cambridge/New York: Cambridge University Press.

Bonk, W J and Ockey, G J (2003) A many-facet Rasch analysis of the second language group oral discussion task, *Language Testing* 20 (1), 89–110.

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

Brown, A (2006) Candidate discourse in the revised IELTS Speaking Test, *IELTS Research Report* 6, 71–89.

Bygate, M (1987) *Speaking,* Oxford: Oxford University Press.

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

Dunlea, J (2015) *Validating a set of Japanese EFL proficiency tests: demonstrating locally designed tests meet international standards*, unpublished PhD thesis, University of Bedfordshire.

Dunlea, J and Figueras, N (2012) Replicating results from a CEFR test comparison project across continents, in Tsagari, D and Csepes, I (Ed) *Collaboration in Language Testing and Assessment,* New York: Peter Lang, 31–45.

Dunlea, J, Fouts, T, Joyce, D and Nakamura, K (2019) EIKEN and TEAP: How two test systems in Japan have responded to different local needs in the same context, in Su, L I-W, Weir, C J and Wu, J R W (Eds) *English Language*

Proficiency Testing in Asia: A New Paradigm Bridging Global and Local Contexts, London/New York: Routledge, 131–161.
Eiken Foundation (2018) *第9回TEAP連絡協議会レポート [The report of the 9th TEAP Liaison Council],* available online: www.eiken.or.jp/teap/group/
Eiken Foundation (no date) *平成31年度事業計画書 [Annual organisation operational plans April 2019–March 2020]*, available online: www.eiken.or.jp/association/report/
Field, J (2011) Cognitive validity, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Second Language Speaking,* Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press, 65–111.
Galaczi, E D, ffrench, A, Hubbard, C and Green, A (2011) Developing assessment scales for large-scale speaking tests: a multiple-method approach, *Assessment in Education: Principles, Policy & Practice* 18 (3), 217–237.
Green, A (2014) *The Test of English for Academic Purposes (TEAP) Impact Study: Report 1 – Preliminary Questionnaires to Japanese High School Students and Teachers*, available online: www.eiken.or.jp/teap/group/pdf/teap_washback_study.pdf
Inoue, C (2013) *Task Equivalence in Speaking Tests: Investigating the Difficulty of Two Spoken Narrative Tasks*, Frankfurt am Main: Peter Lang.
Iwashita, N, Brown, A, McNamara, T and O'Hagan, S (2008) Assessed levels of second language speaking proficiency: How distinct?, *Applied Linguistics* 29 (1), 24–29.
Kormos, J and Dénes, M (2004) Exploring measures and perceptions of fluency in the speech of second language learners, *System* 32, 145–164.
Linacre, M (2011) *Facets Computer Program for Many-facet Rasch Measurement,* Beaverton: Winsteps.com.
McNamara, T F (1996) *Measuring Second Language Performance,* Harlow: Longman.
Messick, S (1996) Validity and washback in language testing, *Language Testing* 13, 241–256.
MEXT (2009) *The course of study for upper secondary school,* available online: www.mext.go.jp/a_menu/shotou/new-cs/index.htm
Nakatsuhara, F (2010) *A background review report: The development of the Test of English for Academic Purposes (TEAP) speaking paper for Japanese university entrants,* project report submitted to Eiken Foundation of Japan.
Nakatsuhara, F (2014) *A research report on the development of the Test of English for Academic Purposes (TEAP) Speaking Test for Japanese university entrants – Study 1 & Study 2,* available online: www.eiken.or.jp/teap/group/pdf/teap_speaking_report1.pdf
Nakatsuhara, F (2018) Investigating examiner interventions in relation to the listening demands they make on candidates in oral interview tests, in Wagner, E and Ockey, G (Eds) *Assessing L2 Listening: Moving Towards Authenticity,* Amsterdam: John Benjamins, 206–225.
O'Sullivan, B and Weir, C J (2011) Test development and validation, in O'Sullivan, B (Ed) *Language Testing: Theories and Practices*, Basingstoke: Palgrave Macmillan, 13–32.
O'Sullivan, B, Weir, C J and Saville, N (2002) Using observation checklists to validate speaking-test tasks, *Language Testing* 19 (1), 33–56.
Pollitt, A and Murray, N L (1996) What raters really pay attention to, in Milanovic, M and Saville, N (Eds) *Performance Testing, Cognition and Assessment: Selected Papers from the 15th Language Testing Research*

Colloquium, Studies in Language Testing volume 3, Cambridge: UCLES/ Cambridge University Press, 74–91.
Sasaki, M (2008) The 150-year history of English language assessment in Japanese education, *Language Testing* 25 (1), 63–83.
Tavakoli, P and Foster, P (2008) Task design and second language performance: The effect of narrative type on learner output, *Language Learning* 58 (2), 439–473.
Taylor, L (2011) Introduction, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Second Language Speaking,* Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press, 1–35.
Taylor, L (Ed) (2011) *Examining Speaking: Research and Practice in Assessing Second Language Speaking,* Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press.
Taylor, L (2014) *A report on the review of test specifications for the Reading and Listening papers of the Test of English for Academic Purposes (TEAP) for Japanese university entrants*, available online: www.eiken.or.jp/teap/group/pdf/teap_rlspecreview_report.pdf
Weir, C J (1983) *Identifying the language problems of the overseas students in tertiary education in the United Kingdom,* unpublished PhD thesis, University of London.
Weir, C J (2005) *Language Testing and Validation: An Evidence-based Approach,* Basingstoke: Palgrave Macmillan.
Weir, C J (2014) *A research report on the development of the Test of English for Academic Purposes (TEAP) Writing Test for Japanese university entrants*, available online: www.eiken.or.jp/teap/group/pdf/teap_writing_report.pdf
Wigglesworth, G and Elder, C (2010) An investigation of the effectiveness and validity of planning time in speaking test tasks, *Language Assessment Quarterly* 7, 1–24.
Wright, B and Linacre, M (1994) *Reasonable mean-square fit values,* available online: www.rasch.org

7 Testing practices and construct operationalisation: reflections on Cyril Weir's view of integrated assessment tasks

Guoxing Yu
University of Bristol
Tony Clark
Cambridge Assessment English

This chapter presents our review and reflections on Professor Cyril Weir's thinking concerning the use of integrated reading/writing tasks (especially summary writing) at different stages of his academic career, from his PhD dissertation (1983) to his most recent publications in the Studies in Language Testing (SiLT) series. Our analysis of his writings as well as his fellow researchers' comments on his publications showed that he was initially very positive about integrated reading/writing tasks in the early 1980s as evidenced in his PhD dissertation, though he did have some reservations due to the challenges in scoring reliability. However, he became more concerned about the measurement 'muddiedness' of such tasks during the 1990s. Towards the end of the first decade of the 21st century, he gradually changed his views, and became again more receptive, optimistic, and eventually advocated enthusiastically in favour of integrated writing tasks. From our analysis we can see the evolution of his views in the last 35 years on integrated reading/writing tasks, especially in the context of English for Academic Purposes (EAP) assessment.

Introduction

I (first author, Yu) still remember the Language Testing Forum I attended in 2002 at the University of Reading. At the conference, I presented my PhD research design – *Reading for Summarization to Measure Reading Comprehension Abilities: Promises and Problems* – as a poster. The five poster presenters were given 5 minutes each to talk about their research to the whole group of conference attendees, without any visual aid or PowerPoint slides. After the oral presentations, the group were given ample time to talk to the presenters at their poster stand. As my PhD research project (Yu 2005) was inspired by Cyril Weir's concern over 'muddied measurement' (Weir 1993)

and Lynda Taylor's PhD dissertation on text-removed summary completion tasks (Taylor 1996, 2013), I was delighted that Cyril came to my poster stand. We had an engaged discussion on a range of issues from my research design to his concept of 'muddied measurement' and my counterargument of 'organic assessment'. As I had anticipated from reading Weir (1993), he argued that summarisation is a 'muddied measurement' because it would involve both reading and writing and that it was not at all clear how precisely each skill would contribute to the successful performance of summarisation. He also reiterated the challenges in marking summaries reliably and consistently. To defend myself, I presented my concept of 'organic assessment'. I put forward an analogy: organic and muddy but tasty carrots from a local organic green grocery compared against those shiny but tasteless carrots from a supermarket. I argued that although the organic carrots may look muddy, we are able to wash off the mud and still get the tasty carrots; while the supermarket carrots may look shiny and beautiful, there is no way that we can make them as tasty as the organic carrots. Similarly, I argued, test takers' performance should be elicited from a natural and organic setting, even though it may look muddied in the first place, so that language testing researchers/providers can extract and infer as much as possible about test takers' language abilities, through statistical analysis or other means, from the muddied but organic assessment data. I argued that it would be counter-productive if we attempted to purify our assessment tasks because such tasks would become too artificial, non-authentic, and not 'organic'. Cyril did look slightly convinced by my analogy; and encouraged me to take a holistic view to explore simultaneously multiple factors of summarisation tasks. Following Cyril's advice and under the expert supervision of Pauline Rea-Dickins, I further refined my research design to make it more multi-faceted. In the dissertation project, I investigated how various factors such as features of source texts (length, summarisability, presentation modes), use of different languages (English, Chinese) and different scoring criteria might all affect students' performance in summary writing.

Cyril's concern over 'muddied measurement' might have been initiated by his research on the Test in English for Academic Purposes (TEAP), as part of his PhD (Weir 1983). The area of study chosen for his PhD, awarded at the Institute of Education, was considerably ahead of its time. By 'identifying the language problems of the overseas students in tertiary education in the United Kingdom' (Weir 1983), he had sought to understand a complex issue, one that is even more salient and germane now than it was then. With the total number of overseas learners enrolled in higher education institutions in the UK expanding from 29,560 in 1978 (Weir 1983:14) to 450,660 in 2017 (UK Council for International Student Affairs, UKCISA), the lines of enquiry, findings and discussion of Cyril's PhD dissertation become more relevant to contemporary research on international students studying in English as a medium of instruction (EMI) universities. Research underpinning the TEAP

test and describing the challenges international students typically encounter – in addition to evaluating the efficacy of various assessment formats (including summary writing) for international in relation to home students – was the principal objective of his dissertation. A methodological framework for doing so was created, designed to encompass a range of levels and diverse subjects, and to contribute to discussions around best practice that continue to this day. (See Chapter 1, this volume, by Vivien Berry on needs analysis with her own evaluation of the contribution of Cyril's early work in this area.)

Numerous studies, including our own (e.g. Rea-Dickins, Kiely and Yu 2007 and Clark 2018), have since investigated the experiences of international students studying in EMI universities from multiple perspectives, e.g. in terms of the challenges international students face in academic writing, the relationships between students' International English Language Testing System (IELTS) test preparation, test scores and their academic engagement and success. The major question that these studies have raised points to a fundamental issue in the assessment of international students' English language proficiency: what to assess and how to assess. Both Rea-Dickins et al (2007) and Clark (2018) have concluded that it would be desirable for IELTS to include reading/writing integrated tasks such as summary writing; however, such recommendations are hardly new or surprising, given that the debates among scholars on whether summary writing should be included or not in large-scale tests to assess international students' English language proficiency have been going on for decades and will continue in the foreseeable future (see also Taylor 2013). In this chapter, we review and reflect on Cyril Weir's thinking concerning the use of reading/writing integrated tasks, especially summary writing, at different stages of his academic career, from his PhD dissertation (1983) to his most recent publications in the SiLT series. Such analysis will enable us to reflect on what we have learned from Cyril Weir and why we are now better for it.

Initial excitement about integrated assessment tasks (1980s)

In his PhD dissertation project[1], Cyril Weir first tried to identify the language problems overseas students encountered in tertiary education in the UK, and then developed some integrated assessment tasks including a summary writing task. Commenting on the potential advantages of using summary writing over other types of tasks, he wrote:

1 Available online: ethos.bl.uk/OrderDetails.do?did=1&uin=uk.bl.ethos.480860. See also www.teachingenglish.org.uk/sites/teacheng/files/ELT-14-screen_0.pdf#page=24

> We viewed summary as potentially **the most valid test** of a student's writing ability in terms of the tasks he has to cope with in the academic situation. The writing of reports and essays at tertiary level requires the ability to select relevant facts from a mass of data and to re-combine these in an acceptable form. Summary of the main points of a text in this fashion involves not only reading and/or listening comprehension, but also the ability to write a controlled composition containing the essential ideas of a piece of writing and omitting non-essentials.
>
> The main difficulty with this component is marking the product reliably and consistently. To evaluate students' responses reliably one needs to formulate the main points contained in the extract, construct an adequate mark scheme and effectively standardise suitable markers to the scheme. Some subjectivity inevitably remains and it is easy to underestimate the difficulty of marking a summary of this type reliably.
> (Weir 1983:377; emphasis added)

Weir's excitement about the promises of using integrated assessment tasks, especially summary writing tasks, was clearly evident throughout his PhD dissertation; however, he remained cautiously optimistic largely due to his concerns about the subjectivity involved in marking summaries and the difficulty in establishing whether the failure of performance in writing was due to 'faulty comprehension of the written text' (1983:347).

> In these integrated tasks we would not be concerned with directly testing the "discrete" enabling skills that we had been able to identify but rather attempting to simulate the types of communicative activity students might encounter in an academic context.
> (Weir 1983:321)

> In integrated formats, where reading and/or listening tasks feed into writing tasks there may be a problem in establishing where the process has broken down. We decided that we would need to assess reading separately as a study mode, as well as combining it with listening/writing activities in order to see if any resultant problems in coping with the integrated task were due to faulty comprehension of the written text.
> (Weir 1983:347)

> There is some evidence in the factor analysis that, owing to the integrated nature of some of the tasks where reading and/or listening feed into writing, performance on the latter is to a certain extent influenced by proficiency in the other skills. The writing scores might, therefore, be **contaminated** by previous performance on listening and/or reading tasks. As these integrated measures reflect the situation students are likely to face in the academic context, we felt that this was acceptable.
> (Weir 1983:552–553; emphasis added)

While acknowledging potential challenges and difficulties in test construction, Cyril Weir believed that such integrated tasks would 'ensure a greater degree of content and face validity for future E.A.P. tests' (Weir 1983:550).

> As with all new departures, integrated, communicative tests are at present difficult to construct, complex to take, difficult to mark and difficult to report results on.
>
> It is felt, however, that the methodological approach we have advocated in this work will help to ensure a greater degree of content and face validity for future E.A.P. tests conceived within this paradigm.
> (Weir 1983:550)

Fellow researchers, especially those who were members of the working party that oversaw the development of the Test in English for Educational Purposes (TEEP), were perhaps equally excited about Weir's methodological approach to research and test development as well as his research findings. For example, Alderson (1988) highlighted that the test was 'developed in a quite different manner'. He went on: 'partly in reaction to the non-empirical way in which the ELTS [English Language Testing Service] was developed, and the associated criticism, Cyril Weir spent two years devising questionnaires and observation schedules and gathering data, under the guidance of a working party. He sought to identify information on the study demands placed on overseas students in various educational settings (university and college) in the UK' (1988:222). Alderson further elaborated that future development of EAP tests does not need to undertake similar empirical needs analysis (see Chapter 1, this volume, by Vivien Berry) of the requirements of tertiary-level students because we can consult and benefit from Weir's 'monumental work' and the 'substantial database' he created for this project.

> This monumental work is available (Weir 1983) for consultation by future test developers, and it represents a major achievement in empirical needs analysis, such that no similar undertaking need be repeated in the foreseeable future for subjects such as Weir's at least, as it provides a substantial database for EAP test development if one is required. The TEEP Project, however, encountered problems in design and execution which were perhaps unavoidable, but which future test developers would do well to pay attention to and benefit from.
> (Alderson 1988:223)

Commenting on Weir's (1983) assessment tasks (which integrated reading and writing, and reading, writing and listening) and his efforts to carefully 'separate out scores of the various skills to avoid what he calls muddied measurement' and the old IELTS test which 'deliberately included input from the reading test in a writing task' (Alderson 1999:64), Alderson (1999) put forward

his usefully provocative 15 dilemmas for reading assessment. Dilemma 5 asked 'can tests test integrated abilities?'. He answered unequivocally – 'the answer to this is clearly yes' – and further asserted that 'there is surely no reason why such integrated tests cannot be developed' because 'insofar as the distinction into four discrete skills is thought to be either invalid, or at least limited or possibly distorting in its view of language use' (1999:65).

Other scholars, for example Alan Davies, were also very positive about Cyril Weir's innovative approach to developing integrated assessment tasks. In his review of the history of IELTS as an example of EAP assessment, Davies (2008) wrote:

> From the 1960s onwards research and development in communicative language testing was much discussed though less often practised. Researchers in Canada (Wesche 1983), in Australia (Keats 1962) and in the UK attempted to marry ideas of performance and authenticity with the constraints of large-scale testing. Most innovative were Morrow (1977), McEldowney (1976) and Weir (1983).
> (Davies 2008:53)

Davies (2008:71) further elaborated on and praised the '*brave* [emphasis added] attempt to develop a communicative test of English for Academic Purposes' initiated by the Associated Examining Board (AEB) and directed by Cyril Weir. The test Cyril Weir developed came to be known as the TEEP, intended for students who would study in EMI academic programmes. The TEEP test became operational in 1984 and is still used by applicants for a degree course at the University of Reading and a number of other UK universities as evidence of their English language skills (see www.reading.ac.uk/ISLI/study-in-the-uk/tests/isli-test-teep.aspx). Davies praised the TEEP test highly. He wrote: 'The TEEP test was *distinct* for two reasons: first, that it was established from the outset as a communicative test and second, that it was planned to provide diagnostic feedback for students and the institutions they were or would be attending' (2008:71; emphasis added). Davies was particularly impressed by the methodological approach to research and development for the new test, which he wrote were 'carefully and deliberately planned in three phases' (2008:71):

> 1. To establish the levels, discipline areas and institutions where overseas students enrol in further and higher education sectors.
> 2. To ascertain the language demands made on students in the disciplines most commonly studied by overseas students.
> 3. To construct a test battery to assess a student's ability in performing the language tasks relevant to the academic context in which they have to operate.
>
> (Davies 2008:71)

In addition to its innovative approach to developing the TEEP test, the findings of Cyril Weir's doctoral dissertation (1983), which reported the research and test development process of TEEP, also influenced in a positive way the major IELTS revision of 1995. Along with findings by colleagues such as Criper and Davies (1988) and Alderson and Urquhart (1985a, 1985b) around the same time, Davies (2008:76) argued that findings of Weir's (1983) research offered 'little support for a test with subject modules', which contributed to the major IELTS revision of 1995. Both Davies (2008:59) and Clapham (1996) quoted Weir's (1983:549–550) concluding comments on the construction of TEEP to account for the decision to remove subject modules from IELTS after its major revision of 1995.

> In our investigations of the language events and activities overseas students have to deal with in British academic environments and the difficulties they encounter therein, we discovered much that was common between students of different disciplines and at different levels. This did not remove the possibility though that the subject content of texts employed in our test tasks might unduly affect performance. Whilst we attempted to take account of this in our sampling, we were unable to produce any conclusive evidence that students were disadvantaged by taking tests in which they had to deal with texts other than those from their own subject area. The case for a variety of ESP tests therefore remains unproven.
> (Weir 1983:549–550)

Clearly these publications (Alderson 1988, 1999, Clapham 1996, Davies 2008) show how much Cyril Weir's research, especially its methodological approach to test development and the findings on the commonalities across different subjects and the enormous benefits of using integrated writing assessment tasks, has been appreciated and has influenced the thinking and practice of EAP assessment in the 1980s–90s.

Concerns about integrated assessment tasks: muddied measurement (1990s to early 2000s)

However, into the 1990s, Cyril himself seemed to become more concerned about integrated writing tasks, so much so that he started to doubt the value of integrated writing assessment tasks. He began to use the term 'muddied measurement', first in his book *Communicative Language Testing* (Weir 1990:85), as well as in several other publications during this period, for example:

> It may be failure in writing and/or reading that has been the stumbling block. If we want to make comments about a student's reading ability per

> se, then this may be taken as an argument for more discrete reading tests if we are to avoid **muddied measurement** from skills integration.
> (Weir 1993:176; emphasis added)

In Urquhart and Weir (1998:121), they suggested avoiding:

> . . . tasks such as selective summary based on prior reading of texts – where the extended writing involved in task completion might interfere with extrapolations we might wish to make concerning candidates' reading abilities alone.

Weir (2005:88) reiterated his concerns regarding 'muddied measurement' throughout his bestselling book on an evidence-based approach to test validation.

> . . . given that in many places in the world employers, admissions officers, teachers and other end-users of test information want to know only about a candidate's reading ability per se, then we must where appropriate address the problems in testing this and try to avoid other constructs, such as writing ability, interfering with its measurement.

When Weir (2005) commented on integrated listening/writing tasks to measure listening comprehension, 'muddied measurement' was again raised as a serious issue:

> In the latter case [testing understanding of a spoken passage through an integrated writing task such as a selective summary of the discourse] the danger of muddied measurement cannot be ignored, i.e., are we testing listening and/or writing?
> (Weir 2005:101)

The above extracts indicate Cyril Weir's apprehension about the risks inherent in failing to isolate one construct of assessment from other overlapping constructs, and the possibility that students may be unfairly penalised if assessment practices are somewhat unrefined or ill-considered. The notion of integrated assessment was not dismissed outright; but Cyril Weir's concerns about the 'muddied measurement' were evident in his publications around that time as shown above. As a task became less direct in nature, so too did the test score it produced become less valuable an indicator of students' reading ability, as it strayed into territory beyond its reasonable jurisdiction. The concept of 'contamination' of scores recurred, suggesting that poorly chosen integrated tasks constructed without due diligence would have such a potentially negative impact on measuring students' performance that it must be treated with a high level of caution. However, even at this early stage, Cyril

Weir noted that this caution may not extend to all contexts, depending on the constructs in question, which in turn depended on the post-test environment that candidates would find themselves in. The academic sphere was singled out by example as one such area in which properly designed integrated skills assessment would be an authentic and therefore desirable reflection of the skills required of the domain (Weir 1983).

Promoting integrated assessment tasks: after 2005

Citing Weir's (1983) findings as supporting evidence, Bruce and Hamp-Lyons (2015:70) wrote:

> The argument that assessment in writing in academic contexts should probe not only the students' general linguistic proficiency but also their ability to handle academic content is a long-standing one.

The aforementioned openness of Cyril Weir to new ideas (or revisiting his previous views, in this case), especially as time passed and mounting evidence in support of an alternative position emerged, appeared to encourage him towards a shift in stance. In later work, Weir's position had shifted slightly, and he mentioned the value of integrated assessment, though always with the caveat that it must be thoroughly underpinned by rigorous research and validation activities. Below we report our analysis of Weir's publications in the SiLT series.

In SiLT Volume 26, *Examining Writing* (Shaw and Weir 2007), Weir seemed to be a lot more consonant or at ease with reading-into-writing tasks like the CAE (Certificate in Advanced English, now known as C1 Advanced) Part 1 task. The authors recognised that 'such reading-into-writing activities are well supported in the current research literature … and are increasingly used in high-stakes Writing tests around the world, for example, in new TOEFL and since the 1980s in TEEP' (Shaw and Weir 2007:74). When they talked about response format as part of 'context validity', they seemed unsatisfied with the limited use of integrated reading and writing tasks in CAE and CPE (Certificate of Proficiency in English, now known as C2 Proficiency), which are recognised for admission purposes in the UK universities. They wrote:

> Another issue for attention is the role of integrated Reading and Writing tasks. CAE and CPE are recognised for university entrance purposes in the UK but in their present format only include tasks which integrate reading and writing in a limited way; such tasks would better reflect reading to learn and writing in that target discourse community and

> are more likely to activate knowledge transformation which, as we have already seen, is the **hall mark of writing at this level** . . .
> (Shaw and Weir 2007:246 (emphasis added), see also Weir, Vidaković and Galaczi 2013:251, 435 for a similar statement.)

However, they rightly expressed their concerns about 'lifting' from source texts and how to deal with that in rating criteria. They also called for further research on integrated reading and writing tasks so as to promote the potential positive washback of such tasks (see also Cumming 2013, Yu 2013a on the promises and perils of using integrated writing tasks, and future research agenda).

> Integrated tasks are not without their disadvantages however, not least in how to deal with candidates "lifting" from the input texts provided; ways will have to be sought to eliminate this in preparing candidates for such an examination task. Punitive sanctions might also be considered to discourage "lifting", e.g. candidates will be penalised if more than X number of continuous words are lifted from the source text(s). The whole area of integrating reading and writing activities is in need of further research but the potential positive washback of such integrated tasks should encourage further research of this nature . . .
> (Shaw and Weir 2007:247, see also Weir, Vidaković et al 2013:435 for a similar statement.)

In SiLT volume 29, *Examining Reading* (Khalifa and Weir 2009), Weir re-emphasised the rationale for promoting reading-into-writing tasks as Shaw and Weir (2007:74) did earlier, and further elaborated on the rationale and reaffirmed the value of summary or an integrated reading-into-writing activity, with greater enthusiasm.

> There is obviously a good case for providing input in writing tests where provision of stimulus texts reflects the real-life situation (e.g. in response to an informal email from a friend at the lower levels, or the writing of university assignments at the higher levels). The highest level of processing in our model discussed in Chapter 3 is where students have to integrate information across texts to develop a combined representation of the texts they have read. . . . Summary or an integrated reading-into-writing activity would seem to be the most appropriate techniques for doing this. Such an approach also helps ensure equal access to domain knowledge among candidates and reduces the potential bias that such internal knowledge can have.
> (Khalifa and Weir 2009:90)

As in Shaw and Weir (2007), Khalifa and Weir (2009) revisited the issues of lifting and the implications for the development and implementation

of rating scale. Khalifa and Weir (2009) made some suggestions on how to deal with borrowing, lifting and plagiarism (see also Weir, Vidaković et al 2013:170).

> Integrating reading with writing activities not surprisingly presents problems for markers in making decisions about what level of borrowing from these texts is permissible and in being confident about what the candidate is capable of actually producing rather than just copying.
>
> The extent of borrowing can be reduced by ensuring that the writing task demands a significant level of input language transformation from the candidate, i.e. the candidate has to do something more than simply lift input material. Additionally, it may be necessary to make clear to candidates what is not permissible in terms of borrowing from text provided and also limits may have to be set on how much text can be quoted as in real-life rules concerned with plagiarism.
> (Khalifa and Weir 2009:91)

As a trained historian, Weir presented an amazingly detailed review of the history of Cambridge English language examinations (1913–2012) in his book *Measured Constructs* (Weir, Vidaković et al 2013). In this book, Weir presented the history of the evolution of summary writing in Cambridge English language examinations, which to a great extent, in our view, might have consolidated his initial excitement for using integrated tasks that he pioneered in his PhD research (1983) and ultimately have further convinced him to advocate integrated assessment tasks enthusiastically. Weir found that 'the integrative approach, in the form of summary, translation, reading aloud and dictation, had been present in Cambridge examinations since 1913' (Weir, Vidaković et al 2013:70). The volume further noted 'how reading was never tested in a separate paper by Cambridge until 1975, but was rather a component of a number of integrated tasks that were favoured at the time. . . . Such integrated tasks would *involve* rather than uniquely *focus* on reading ability' (Weir, Vidaković et al 2013:104; emphases in original).

Whilst acknowledging that summarisation of main ideas at the text level is one of the more demanding levels of processing activity in real-life language use and thereby appropriate only for advanced-level test takers, Weir seemed to be convinced by my rationale for using summarisation as a measure of reading comprehension as well as my research findings (Yu 2008). He wrote:

> Yu (2008:522–23) offers an **impressive** list of references in support of the use of summary in teaching and testing reading and provides empirical support for the use of summary as a test of reading comprehension.
>
> Reading comprehension was the only statistically significant predictor for both English and Chinese summarisation performances.

> Students with better reading comprehension produced better summaries (Yu 2008:544).
> (Weir, Vidaković et al 2013:135 (emphasis added), see also Weir 2014:6–7 for his report on TEAP (Test of English for Academic Purposes) Writing for Eiken Foundation Japan[2] for a similar statement.)

Weir also seemed to become less concerned about the potential problems of marking reliability he had observed in his PhD study (1983), perhaps mulling over the pros and cons of summarisation tasks with the mounting evidence over the previous 30 years.

> Most recently Yu (2008:547) in his doctoral research on summarisation in English and Chinese concluded that "complexities in judging the quality of summarization performance resonate with Weir's (1993:154) concern regarding the subjectivity of marking written summaries". Weir might argue now of course that if we can mark essays reliably, this must be possible for summary too.
> (Weir, Vidaković et al 2013:135)

Weir's enthusiasm for summary writing was abundant, as he continued writing about his regret concerning the demise of summary in CPE in 1975 and his anticipation of the reintroduction of summary writing[3] to CPE in 2012 as Khalifa and Weir recommended (2009). He wrote:

> Summary was to last as a task in CPE right through to 1975 and, given the critical use of CPE for university entrance in the 21st century, the demise of such an authentic academic reading-into-writing task might, **with the advantage of hindsight, be regretted** (note however its return to favour in 2003 albeit in a reduced intertextual form in the Use of English paper) . . . From a present day perspective we would argue that (albeit in an integrated format) summary effectively tests the important advanced level reading skill of creating a text level representation . . . a vital element of academic study, in an authentic manner. **No other task type has filled this vacuum** and the recommendation that summary should be reintroduced in CPE made in Khalifa and Weir (2009:220) will take place in the 2012 version of the writing paper in this examination (the intertextual summary from the Use of English paper is being moved there).
> (Weir, Vidaković et al 2013:137; emphases added)

When talking about the case for providing input in writing tests where provision of stimulus texts reflects the real-life situation, he was quite assertive

2 Available online: www.eiken.or.jp/teap/group/pdf/teap_writing_report.pdf
3 See www.cambridgeenglish.org/exams-and-tests/proficiency/exam-format/

in his evaluation of summary or an integrated reading-into-writing activity. He wrote: 'Summary or an integrated reading-into-writing activity would seem to be among the **most appropriate** techniques for doing this. Such an approach also helps ensure equal access' (Weir, Vidaković et al 2013:169; emphasis added). His positive evaluation of summary writing continued, arguing that summary writing is capable of promoting the involvement of knowledge transformation.

> By adding the additional summary task this broadened the base for evaluating the candidate's competence in writing and addressed a number of the concerns we raised above about the single essay format for writing. It also meant the process of writing was, through the summary task, more likely to have involved knowledge transformation as well as knowledge telling . . .
> (Weir, Vidaković et al 2013:225)

In addition to the SiLT publications, we also analysed the other publications written by research teams at the University of Bedfordshire led by Cyril Weir. For example, from the research reports to Eiken Foundation of Japan on its TEAP[4] (Test of English for Academic Purposes) integrated writing tasks (Weir 2014), and to the Language Training and Testing Center (LTTC) in Taiwan on its General English Proficiency Test (GEPT) Advanced Writing Task 1 (Chan, Wu and Weir 2014:76), and Chan's (2013) PhD under Weir's supervision, we can see how Weir and his colleagues are actively promoting and researching integrated reading-into-writing tasks (see also Chan and Latimer, Chapter 5, this volume). The concession that integrated assessment, in its contemporary form at least, has the potential to foretell candidates' subsequent academic writing capacity was an open-minded and evidence-based move towards testing which 'embraces both constructs' (Weir, Chan and Nakatsuhara 2013:22), reading and writing in the case in question.

Conclusion

Our analysis of Cyril Weir's publications showed that he was initially very positive about integrated reading/writing tasks in the early 1980s as evidenced in his PhD dissertation, though with some reservation due to the challenges in scoring reliability. However, he became more concerned about the muddiedness of measurement of such tasks in the 1990s. Towards the end of the first decade of 2000, he gradually changed his views, and became

4 This is distinct from the test designed originally by Cyril Weir in the 1980s named TEAP (Test *in* English for Academic Purposes) which was then was changed to TEEP (still used by the University of Reading).

again more receptive, optimistic, and eventually advocated enthusiastically in favour of integrated writing tasks. From our analysis we can see the evolution of his views in the last 35 years on reading/writing integrated tasks, especially in the context of EAP assessment. The evolution of his views reflects the attributes of a great scholar – his willingness and openness to accept new research evidence. We think this is one example of how he practised firmly and consistently his belief in an evidence-based approach to language test validation. Such analysis also enabled us to reflect on what we have learned from Cyril Weir and why we as a field of language testing are now in a better position to understand the promises and challenges of using summarisation and other types of integrated reading/writing tasks as a measure of writing abilities in EAP contexts, and to explore further the issues that he has identified. The impact of his research on integrated writing assessment tasks in large-scale tests on the field and on individual researchers like ourselves and his colleagues at the Centre for Research in English Language Learning and Assessment (CRELLA) is highly visible now and will be long-lasting.

When he was interviewed for the book *Cambridge English Exams: The First Hundred Years* (Hawkey and Milanovic 2013), Cyril Weir predicted that the next 10 to 20 years would 'see greater attention paid to cognitive processes, so that we might make performance in the test tasks resemble more closely language activities in real life', to make an assessment activity have 'far greater generalisability across a whole range of situations' (2013:342). Summarisation tasks are widely accepted as resembling language activities in real life. They are used in several large-scale English language tests such as Internet-based Test of English as a Foreign Language (TOEFL iBT), Pearson Test of English: Academic, C2 Proficiency, and China's national university entrance exams in English. Taylor (2013:59) and Yu (2013b) presented a comprehensive review of different types of summarisation tasks used in large-scale international English language tests. It is possible that the cognitive processes involved in completing different summarisation tasks may well differ across tasks. Following Cyril Weir's prediction, we would argue that more research efforts now should be put into understanding better the cognitive validity of such summarisation tasks in assessment contexts. As summarisation is central to the successful completion of all kinds of integrated assessment tasks (e.g. reading/writing, listening/reading/writing, and reading/writing/speaking, see Yu 2013a), we anticipate a much broadened and joined-up research effort in the foreseeable future to investigate the extent to which 'summarisation' can play a role in different kinds of integrated assessment tasks, including writing and speaking, whether in a first or a second language (e.g. Zhu, Li, Yu, Cheong and Liao 2016). It is evident that Cyril Weir was pragmatic enough to modify his views as our knowledge of integrated skills assessment and its implications for testing increased, a valuable lesson he has

taught us on the need for academic openness and flexibility, if progress is to be encouraged.

References

Alderson, J C (1988) New procedures for validating proficiency tests of ESP? Theory and practice, *Language Testing* 5 (2), 220–232.

Alderson, J C (1999) Reading constructs and reading assessment, in Chalhoub-Deville, M (Ed) *Issues in Computer-adaptive Testing of Reading Proficiency*, Studies in Language Testing volume 10, Cambridge: UCLES/Cambridge University Press, 49–70.

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

Alderson, J C and Urquhart, A H (1985b). This test is unfair: I'm not an economist, in Hauptman, P C, LeBlanc, R and Wesche, M B (Eds) *Second Language Performance Testing*. Ottawa: University of Ottawa Press, 15–24.

Bruce, E and Hamp-Lyons, L (2015) Opposing tensions of local and international standards for EAP writing programmes: Who are we assessing for?, *Journal of English for Academic Purposes* volume 18, 64–77.

Chan, S H C (2013) *Establishing the validity of reading-into-writing test tasks for the UK academic context,* unpublished PhD thesis, University of Bedfordshire.

Chan, S H C, Wu, R Y-F and Weir, C J (2014) *Examining the context and cognitive validity of the GEPT Advanced Writing Task 1: A comparison with real-life academic writing tasks*, available online: www.lttc.ntu.edu.tw/lttc-gept-grants/RReport/RG03.pdf

Clapham, C M (1996) *The Development of IELTS: A Study of the Effect of Background Knowledge on Reading Comprehension*, Studies in Language Testing volume 4, Cambridge: UCLES/Cambridge University Press.

Clark, T (2018) *Bridging the gap: The relationship between intensive IELTS writing preparation in China and Japan and 'relearning' academic conventions*, unpublished PhD thesis, University of Bristol.

Criper, C and Davies, A (1988) *ELTS Validation Project Report*, London/Cambridge: British Council/UCLES internal report.

Cumming, A (2013) Assessing integrated writing tasks for academic purposes: Promises and perils, *Language Assessment Quarterly* 10 (1), 1–8.

Davies, A (2008) *Assessing Academic English: Testing English Proficiency, 1950–1989 – The IELTS Solution*, Studies in Language Testing volume 23, Cambridge: UCLES/Cambridge University Press.

Hawkey, R and Milanovic, M (2013) *Cambridge English Exams: The First Hundred Years. A History of English Language Assessment from the University of Cambridge 1913–2013*, Studies in Language Testing volume 38, Cambridge: UCLES/Cambridge University Press.

Khalifa, H and Weir, C J (2009) *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing volume 29, Cambridge: UCLES/Cambridge University Press.

Rea-Dickins, P M, Kiely, R and Yu, G (2007) *IELTS Research Reports 7: Student Identity, Learning and Progression: The Affective and Academic Impact of IELTS on 'Successful' Candidates*, Cambridge: British Council/UCLES/IDP.

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

Taylor, L (1996) *An investigation of text-removed summary completion as a means of assessing reading comprehension*, unpublished PhD thesis, University of Cambridge.

Taylor, L (2013) *Testing Reading Through Summary: Investigating Summary Completion Tasks for Assessing Reading Comprehension Ability*, Studies in Language Testing volume 39, Cambridge: UCLES/Cambridge University Press.

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

Weir, C J (1983) *Identifying the language problems of the overseas students in tertiary education in the United Kingdom*, unpublished PhD thesis, University of London.

Weir, C J (1990) *Communicative Language Testing*, Hemel Hempstead: Prentice Hall.

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

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

Weir, C J (2014) *A Research Report on the Development of the Test of English for Academic Purposes (TEAP) Writing Test for Japanese University Entrants*, available online: www.eiken.or.jp/teap/group/pdf/teap_writing_report.pdf

Weir, C J, Chan, S H C and Nakatsuhara, F (2013) *Examining the Criterion-Related Validity of the GEPT Advanced Reading and Writing Tests: Comparing GEPT with IELTS and Real-Life Academic Performance*, available online: www.lttc.ntu.edu.tw/lttc-gept-grants/RReport/RG01.pdf

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

Yu, G (2005) *Towards a model of using summarization tasks as a measure of reading comprehension*, unpublished PhD thesis, University of Bristol.

Yu, G (2008) Reading to summarize in English and Chinese: A tale of two languages?, *Language Testing* 25 (4), 521–551.

Yu, G (2013a) From integrative to integrated language assessment: Are we there yet?, *Language Assessment Quarterly* 10 (1), 110–114.

Yu, G (2013b) The use of summarization tasks: Some lexical and conceptual analyses, *Language Assessment Quarterly* 10 (1), 96–109.

Zhu, X, Li, X, Yu, G, Cheong, C M and Liao, X (2016) Exploring the relationships between independent listening and listening-reading-writing tasks in Chinese language testing: Toward a better understanding of the construct underlying integrated writing tasks, *Language Assessment Quarterly* 13 (3), 167–185.

8 The role of academic institutions in language testing research and consultancy

Lynda Taylor

Centre for Research in English Language Learning and Assessment, Bedfordshire

Anthony Green

Centre for Research in English Language Learning and Assessment, Bedfordshire

This chapter reflects upon the role of academic institutions in language testing research and consultancy over the past 50 years, specifically the part played by university-based departments or research centres in developing theory and practice in the field of language testing and assessment (LTA). It acknowledges a part of the story of our professional field which has not received much attention and it offers an appreciation of the contribution of selected individuals, teams and organisations within that story. The chapter:

- describes a range of academic institutional contexts in which LTA research has flourished at different times over recent decades
- considers who the key players were and what the individuals and teams within them accomplished
- reflects upon the significance and impact of such institutions in the field and their legacy with regard to current theory and practice in LTA.

Introduction

In 2005, Professor Cyril J Weir was appointed as the first Director of the Centre for Research in English Language Learning and Assessment (CRELLA), a self-funding research facility within the University of Bedfordshire, UK. Cyril conceived CRELLA as a UK-based centre that would provide quality research and development in the areas of English language learning and assessment. Over the following decade, Cyril and his growing team of colleagues succeeded in building CRELLA to become a centre of excellence that won both national and international recognition. Today, the centre continues to provide examining boards and government organisations with consultancy in matters of test design, development and review, as well as opportunities for

PhD students, post-doctoral fellows and experienced academics to engage in language testing research.

CRELLA is a recent and particularly successful example of a language testing research centre established within a university-based context, but it is not unique. Over the past 50 years, a number of other departments or centres have developed at different points in time and in different locations, mainly, though not exclusively, in the UK, North America and Australasia. Some of these have enjoyed considerable longevity and continue to be active and evolve, sometimes reinventing or reinvigorating themselves. Others enjoyed a shorter lifespan, sometimes morphing over time into a different form or merging with other parts of the parent institution.

This chapter considers the role and contribution of different language testing centres located within an academic institution whose remit is, or has been, to conduct research into language testing and assessment for the benefit of the profession, including practitioners in institutional testing units and within national and international examination boards. In some cases, a dedicated centre was intentionally established for the purposes of language test development and research. In other cases, a particular university department or faculty developed for itself a strong language testing and assessment focus, often as a result of the interest and expertise of a staff member, leading to important research and test development activity. In one case the story is one of collaboration *across* different academic institutions for the simple reason that language testing specialists are few and far between, often located within applied linguistics-related or general education faculty departments.

In homage to Cyril Weir's dedication towards ensuring that the history of our field is properly chronicled, our aim in this chapter is to chart the development of some of these centres, exploring why and when they came into being, who the key players were, and what the individuals and teams within them accomplished.

Background to the chapter

Gathering information to explore the history of each university-based context discussed below proved more challenging than we first anticipated. As Cyril himself often observed, institutions – whether large or small, academic or otherwise – do not always do a good job of recording their origins or chronicling their story as they develop over time. 'Institutional memory' is not always preserved by successive generations and is rarely treasured by the institution itself. This proved to be the case for several of the university-based departments or centres that we listed when we began to scope this paper. Information about when and how they were set up, who was involved, how they developed, what they achieved and what happened to them over time had to be gleaned piecemeal from various sources. Sometimes these sources

were more formal, e.g. desk research on academic publications and across university websites. Sometimes we resorted to informal sources, e.g. postings to the professional listserv for language testers and email requests to specific individuals. The CVs of language testing professionals, together with their interviews in journal articles, proved to be a rich source of information on the impact and legacy of some institutions, as did other publications that explicitly set out to chronicle the history of our field. Although some of what we write has an anecdotal quality to it, we hope that this will make our account more engaging and, in some ways, may allow professional colleagues to tell their own story.

Given space constraints, it has not been possible to do justice to all the university-based research centres and academic departments which could be considered to have made a contribution to the field of LTA research and consultancy over the past half century. Our selection is subjective and our coverage can only be partial. We have included key institutions that occurred to us, particularly those with which Cyril himself had personal involvement, and we can only apologise for significant omissions from the list in the eyes of others. We have chosen to explore the story of these institutions as a means of illustrating when, where and how research centres such as these were active and as a way of reflecting upon their key achievements and long-term impact on the field. For reasons of space, we have limited ourselves to academic contexts that are primarily English-speaking, with the exception of Canada where bilingualism is an important consideration in education and wider society. However, we recognise the important research contribution made, especially in recent years, by university-based language testers and centres in other linguistic contexts in the Far East (e.g. Japan, China, Hong Kong, Taiwan and Korea) and in Europe (e.g. Finland, Italy, Germany), as well as in other parts of the world such as Israel and South Africa. Hopefully, their stories can be gathered in due course and added to those recorded here.

We should also acknowledge that important research facilities and activities located within examination boards, even if associated with a university institution (such as Cambridge in the UK or Michigan in the US), will not be considered here since these are somewhat different in nature. Furthermore, the history and legacy of some of these organisations are already well documented. Spolsky (1995) and others have dedicated many pages to recording the story of Educational Testing Service (ETS), while the story of UCLES/Cambridge ESOL (now Cambridge Assessment English, henceforth Cambridge English) has been chronicled in depth by Cyril himself and by others in numerous volumes in the Studies in Language Testing (SiLT) series (Hawkey and Milanovic 2013, Weir and Milanovic (Eds) 2003, Weir, Vidaković and Galaczi 2013).

Structure of the chapter

The contexts we have chosen to focus on are:

- Department of Applied Linguistics, University of Edinburgh (UK)
- Department of Applied Linguistics, University of California, Los Angeles (UCLA) (US)
- Testing and Evaluation Unit, University of Reading (UK)
- Department of Linguistics and English Language, Lancaster University (UK)
- Language Testing Research Centre (LTRC), University of Melbourne (Australia)
- Centre for Research in Testing, Evaluation and Curriculum (CRTEC), University of Roehampton, Surrey (UK)
- The Canadian 'academic network': university-based institutions in Montreal, Ottawa, Toronto, British Columbia, etc.
- Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire (UK).

A brief description of some key aspects of the life and legacy of each of these academic contexts is presented below. Institutions are roughly sequenced according to when they were first established or most active in our view – rather than in any order of importance or seniority! Later in the chapter we reflect on some common and recurring themes that emerge from their respective stories, including their impact upon: the evolution of theory and practice in LTA; the growth of a professional cadre and community of well-qualified language testers; the expansion of knowledge and good practice through publications, conferences, training courses, etc.; and the applied research and development underpinning language tests, testing systems and other forms of language assessment.

Department of Applied Linguistics, University of Edinburgh

Following the pioneering work of Robert Lado as Director of the English Language Institute at the University of Michigan in the 1950s/60s (Lado 1957, 1961, 1964), the Department of Applied Linguistics at the University of Edinburgh in the late 1960s and early 1970s was perhaps the first university-based department outside the US to dedicate serious time and energy to the theory and practice of researching and designing language tests. Out of the work of the 'Edinburgh school' of applied linguists, which included S Pit Corder, J P B Allen, H G Widdowson, Anthony Howatt, Julian Dakin and Gillian Brown, emerged a strong interest in the interface between applied linguistics, language pedagogy and language testing/assessment.

Drawing upon their academic teaching programme, members of the Department of Applied Linguistics prepared an integrated series of textbooks for students consisting of material selected, developed and tested within the department. The series was entitled the *Edinburgh Course in Applied Linguistics* (ECAL) published by Oxford University Press. Volume 3 was *Techniques in Applied Linguistics* (Allen and Pit Corder (Eds) 1974) and it examined a wide range of techniques involved in the planning of language teaching and the preparation of teaching materials. The book included a 30-page section by Elisabeth Ingram addressing the following topics: *definition of a test*; *requirements of a test*; *types of test item*; *the testing of tests*; *language testing exercises*; and *practical work*. Further reading was limited to just four authors at that stage: Davies (1968), Lado (1961), Valette (1968) and Vernon (1956).

By 1977 a fourth volume had been added to the ECAL series, edited by J P B Allen and Alan Davies and entitled *Testing and Experimental Methods* (Allen and Davies (Eds) 1977). This new volume dedicated 10 sections (233 pages) – as compared to the single section in Volume 3 – to matters of testing, assessment and research design. The publication of Volume 4 five years after the previous volume might suggest that interest in research associated with testing linguistic knowledge and language skills was growing apace, attracting greater attention from university-based applied linguists and language practitioners.

Alan Davies and Elisabeth Ingram were among the earliest applied linguists within a UK university context to focus close attention on the theory and practice of testing language skills and they were well qualified to do so. During the early 1960s Elisabeth Ingram had been instrumental in developing the English Language Battery (ELBA) at the University of Edinburgh in response to the need to assess international students' English especially at postgraduate level. In 1963–65 Alan Davies, as Senior Research Associate, had been involved in the development of the English Proficiency Test Battery (EPTB), a collaborative project between the University of Birmingham and the British Council to create an English language proficiency test for use in the English higher education sector (Davies 2008). His move to Edinburgh began a strong tradition of language testing research and development there which endured until the end of the 20th century. Not surprisingly, the Department of Applied Linguistics at the University of Edinburgh provided the training ground for numerous language testing experts of subsequent decades, including Charles Alderson, Liz Hamp-Lyons, Dan Douglas, Cyril Weir and Neil Jones, among others.

Alan Davies himself went on to serve the wider field in many different ways, including as chair and committee member of the British Association for Applied Linguistics (BAAL), head of the Department of Applied Linguistics at Edinburgh, editor of the journal *Applied Linguistics*, and

Secretary-General of the International Association of Applied Linguistics (AILA). Brumfit (2001) commented that for many applied linguists, Alan Davies came to be identified as the major British theorist in the field of language testing. Brumfit also described him as 'a major humanising influence' across the whole discipline of British applied linguistics (2001:2), perhaps because of his enduring interest in and commitment to the field of ethics and its direct relevance to the use (and misuse) of tests and testing practices. Alan went on to serve as President of the International Language Testing Association (ILTA) in 2000 and led the teams that developed ILTA's *Code of Ethics* (2000) and *Guidelines for Practice in English* (2007).

Department of Applied Linguistics, University of California, Los Angeles (UCLA)

In the US, it was probably the University of California, Los Angeles (UCLA) which for many years could boast, like Edinburgh, one of the strongest Applied Linguistics programmes in the country during the 1970/80s. Sara Cushing (now Professor at Georgia State University) recalls that Grant Henning was on the faculty at UCLA when she started her Master's programme in 1986 before he left to join Educational Testing Service (ETS) the following year (personal communication, 2019). Brian Lynch, an early UCLA graduate, was Sara's first mentor and Frances Butler Hinofotis also did early language testing work there. Scholars graduating from UCLA in those early years included: James Dean Brown and Thom Hudson, both of whom went on to academic careers in language assessment with a particular focus on task-based performance assessment and criterion-based assessment at the Second Language Teaching and Curriculum Center of the University of Hawai'i at Manoa (Brown, Hudson, Norris and Bonk 2002, Norris, Brown, Hudson and Yoshioka 1998); Kathi Bailey and Jean Turner, who went on to the Monterey Institute of International Studies (now the Middlebury Institute of International Studies at Monterey); Fred Davidson, who went on to the University of Illinois at Urbana-Champaign; and Hossein Farhady, an Iranian applied linguist who later returned to Iran and would have a significant impact on language testing research and the training of a generation of researchers in that part of the world.

In the late 1980s, the PhD in Applied Linguistics at UCLA was an interdisciplinary programme with required courses, such as phonetics and syntax, offered by the Department of Linguistics. Lyle Bachman arrived at UCLA in 1990, the publication year for his *Fundamental Considerations in Language Testing* volume (Bachman 1990), to strengthen the interest in and focus on language assessment, but there was never an actual centre or organisational structure focused on language testing *per se* at the university. Interdepartmental flexibility seems to have allowed for highly specialised language

testing courses of various sorts, sometimes with very small numbers of students in class, a *modus operandi* that nowadays would be much more difficult to operate and justify economically within postgraduate education in most countries. For example, Sara Cushing recalls taking Lyle's history of language testing course with fellow student Jim Purpura and only one other student. The first PhD student to graduate with Lyle Bachman from UCLA was Antony Kunnan, now Professor at the University of Macau. Antony went on to teach and write extensively on language assessment issues as well as to become founding editor in 2003 of the journal *Language Assessment Quarterly* (LAQ) and founding president in 2014 of the Asian Association for Language Assessment (AALA).

Jim Purpura (now Professor at Teachers College, Columbia University, New York) arrived at UCLA in 1990. Jim recalls taking a number of courses including *Introduction to Language Testing*, through which Lyle and his colleague, Adrian 'Buzz' Palmer, now Associate Professor at the University of Utah, were able to trial material for their seminal volume *Language Testing in Practice*, published by Oxford University Press (Bachman and Palmer 1996) (personal communication, 2019). At Sara and Jim's request, Lyle also offered a course named the *History of Language Testing Research* which provided a comprehensive reading list and later inspired Jim to run a similar course at Teachers College. Other Bachman-initiated courses at UCLA included one on item response theory (IRT, three parameters), a jointly taught course with Tim McNamara on Many-Facet Rasch Measurement and G-Theory, a course on Structural Equation Modelling (SEM) and a course on Criterion-Referenced Assessment. Jim recalls many of these courses as being 'real seminars', where the tutor was learning alongside his students! One course offered each semester was the 'Language Assessment Lab' where students worked as research assistants on grant-funded projects; they also helped to organise conferences and collaborated on conference presentations, journal articles and academic volumes. Later students graduating from the UCLA PhD programme in language testing included Yasuyo Sawaki (now professor at Waseda University in Japan), Xiaoming Xi (now Executive Director, New Product Development, ETS), Gary Ockey (professor at Iowa State University), Nathan Carr (professor at California State University, Fullerton), Lorena Llosa (professor at New York University), Sun Young-Shin (professor at Indiana University) and others who went on to become respected researchers in the field, working in prestigious testing organisations and academic institutions around the world. Jim speaks of being inspired by Lyle's approach to doctoral advisement, in which he not only emphasised scholarship but also service and professionalism. Many UCLA students later went on to be active members of ILTA, including in the role of ILTA President.

Testing and Evaluation Unit, University of Reading

The Testing and Evaluation Unit (TEU) was set up at the University of Reading under the umbrella of the Centre for Applied Language Studies (CALS) established by David Wilkins in 1974. Under Wilkins and his successor, Ron White, CALS was a thriving self-supporting centre that earned its keep from language courses, teacher training and consultancy projects as well as Master's and doctoral programmes. Among the first members of CALS were Keith Morrow and Keith Johnson (the two Keiths as they were known), who did much to translate David Wilkins' proposals for notional/functional syllabuses into the earliest communicative language tests. Arthur Hughes, who taught applied linguistics at the university for 25 years, was prominent in establishing the TEU and directing its activities, working with Tony Woods, who brought essential expertise as a statistician. Activities included designing English language tests for the university and other test providers, offering consultancy services and organising courses and workshops (often funded by the British Council), as well as teaching on postgraduate programmes in applied linguistics and TEFL.

With Don Porter, his successor as director of the TEU, Arthur Hughes set out to build an international network of language testing practitioners and researchers. In 1983, they published *Current Developments in Language Testing*, based on a seminar held at Reading, featuring discussion of the hot topic of the day: John Oller's unitary competence hypothesis (Hughes and Porter (Eds) 1983). The following year, building on the 'Language Testing Newsletter' that Don had established, they launched the first peer-reviewed journal dedicated to our field, *Language Testing,* as 'a forum devoted exclusively to the issues which concern those involved with, or simply interested in, the assessment of language ability in one form or another' (Hughes and Porter 1984:i). In 1986, Cyril Weir arrived at Reading to become the TEU's Director at CALS and he stayed until 2000; together with his Reading colleagues he contributed to the growing number of publications on test theory and development (e.g. Hughes 1989, Weir 1990, 1993).

During the 1970s and 1980s, postgraduate students on the Reading MA in Applied Linguistics, and later the MA TEFL courses (from 1983), benefitted greatly from the combined expertise of the TEU team. Nick Saville recalls how the opportunity to study a module on assessment and to do a full dissertation at Master's level helped launch the career of many who subsequently went on to become influential in assessment (personal communication, 2019). He credits Arthur Hughes and Don Porter as having been instrumental in both him and Mike Milanovic getting started in the field and finding work later on with Cambridge Assessment English. The Reading–Cambridge connection continued for a number of years until the end of the century when Cyril moved on to Roehampton and his colleagues at Reading retired.

As well as offering a specialist Master's course in language testing, Reading produced a list of doctoral students who went on to be influential in our field: Peter Storey (Head of the Centre for Language in Education at The Hong Kong Institute of Education), Hanan Khalifa (Director of Education Transformation and Alliances, Cambridge English), Jenny Bradshaw (Head of International Comparisons at the National Foundation for Educational Research), Barry O'Sullivan (Head of Assessment Research & Development, the British Council), Toshihiko Shiotsu (Kurume University), Masashi Negishi (Professor of Applied Linguistics at Tokyo University of Foreign Studies), Alan Tonkyn (Programme Director of the MA in Applied Linguistics at the University of Reading) and Rita Green (Course Director of Language Testing at Lancaster).

In 1984, Don Porter brought the Test of English for Educational Purposes (TEEP) to Reading University. TEEP (originally known as TEAP, with the A standing for 'Academic') has its roots in the extensive PhD study carried out by Cyril Weir (1983) for the Associated Examining Board (AEB, which later combined with other examination boards to form AQA). Cyril's research, which explored the language problems encountered by international students, responded to a growing need to evaluate the suitability of placing non-native English-speaking students on UK university degree courses. John Slaght, Director of Assessment and Test Development at Reading, notes the ground-breaking nature of Cyril's original study for its time; and the methodology he adopted is described and evaluated in Vivien Berry's contribution to this volume (see Chapter 1).

TEEP has been adapted over the years in response to research and in light of the changing nature of international education in the UK, but remains recognisable as the test that Cyril designed. The test, and its associated programme of research, is now the main focus of the TEU under the stewardship of Bruce Howell and John Slaght. They have developed a new speaking element which assesses both monologic and dialogic-type speaking in a topic-based context and a new undergraduate-level variant. Mainly used with students attending pre-sessional courses in English for Academic Purposes (EAP) at the University of Reading, TEEP is also recognised by other universities. Although negatively affected by the 2010 change in UK visa laws, which included the requirement that approved tests should be available at test centres in more than 40 countries (beyond the resources of a small, university-based centre), some 25 UK universities still accept TEEP for direct entry to their academic courses including the universities of Leicester, Leeds, Glasgow, Manchester and Sheffield.

Department of Linguistics and English Language, Lancaster University

From the early 1980s, the Department of Linguistics and English Language at Lancaster University developed a strong reputation for itself as a centre for training, consultancy and research in language testing and assessment. Key personnel in the early years included Charles Alderson (himself a graduate of Edinburgh), Dianne Wall and Caroline Clapham, among others.

In 1985 Lancaster University began publishing a regular newsletter, *Language Testing Update* (LTU), which over the years was edited by Charles Alderson, Pauline Rea-Dickins, Dianne Wall, Caroline Clapham and Jay Banerjee. When the International Language Testing Association (ILTA) was formally established in 1992, LTU became the official newsletter of the new association. Each issue included news not just from ILTA members but from other language testing associations (in Europe, Japan, Israel, Australia) as well as updates on new test developments and assessment policy-making around the world. This allowed individual language testers and testing groups worldwide to share news of events and initiatives and to showcase their research endeavours to a wider audience. LTU ceased publication in 2004 after providing a valuable service to the worldwide community of language testing researchers for nearly two decades.

PhD training in language testing was ongoing at Lancaster from the 1980s and graduates include many well-known names in our field: Pauline Rea-Dickins, Gary Buck, Glenn Fulcher, Caroline Clapham, Jo Lewkowicz, Yoshinori Watanabe, Dianne Wall, Junko Yamashita, Jay Banerjee, Ali Van Moere, Dina Tsagari, Spiros Papageorgiou and Carol Spöttl, all of whom went on to key positions in university departments or testing organisations worldwide. The early years of this century saw the introduction of Lancaster's Language Testing Summer School (from 2001), followed by an online MA in Language Testing (from 2007), and the range of part-time and distance-based courses steadily expanded enabling growing numbers of students and practitioners from around the world to access the teaching and research training opportunities offered by the university. Although there was never a language testing research centre *per se* at Lancaster University, the Language Testing Research Group (LTRG) has been active since the 1980s, providing a place for academic staff and students to meet, present research and collaborate.

In terms of major research focuses over the 1980s and 1990s, Lancaster was a key site for early studies of test washback and impact (Wall 2005); studies into diagnostic assessment at Lancaster played a key role in the development of DIALANG which continues to this day (Alderson 2005); in relation to the assessment of English for Academic Purposes (EAP) and English for Specific Purposes (ESP), teams at Lancaster were closely involved in the revision of

the original English Language Testing Service (ELTS) and the development of the new International English Language Testing System (IELTS) in 1989 (Alderson and Clapham (Eds) 1992, Clapham 1996, Clapham and Alderson (Eds) 1997); and the university also researched the testing of English for military and aviation contexts during the 1990s and 2000s. More recently, the assessment of listening and reading has become a major focus of attention in the work of Luke Harding and Tineke Brunfaut (e.g. Harding, Alderson and Brunfaut 2015), often employing newer research methodologies such as eye-tracking (e.g. McCray and Brunfaut 2018).

Lancaster's practical concern for sound test design and its commitment to encouraging good test development practice is evidenced in the publication in 1995 of *Language Test Construction and Evaluation*, co-authored by Charles Alderson, Caroline Clapham and Dianne Wall, and published by Cambridge University Press. Together with other practice-oriented books by Arthur Hughes (1989) and Cyril Weir (1990, 1993), this was among the early volumes in the UK to offer language testing practitioners, including language teachers, practical and accessible guidance on how to construct and administer a good language test. Charles Alderson went on to play a key role as co-editor (with Lyle Bachman) of the Press's Cambridge Language Assessment series which sought to bridge research and practice, publishing 10 volumes between 1999 and 2006.

The current team at Lancaster – Tineke Brunfaut, Luke Harding and, since 2017, John Pill – continue to engage in a wide range of research consultancy projects worldwide, with a focus on assessment projects and frameworks as part of national educational reform programmes.

Language Testing Research Centre (LTRC), University of Melbourne

Following its establishment in 1990, with seed funding from Australia's National Languages and Literacy Institute of Australia, the Language Testing Research Centre (LTRC) at the University of Melbourne earned an international reputation for its contributions to language testing, language assessment and language programme evaluation over three decades.

Although located within the University of Melbourne, test development, consultancy and research projects at the LTRC were funded primarily from external sources from the outset (see McNamara 2001 for an account of the Centre's early history). Projects involve a wide array of contexts, including: tests for immigrants; tests for young learners; language tests for specific purposes (e.g. health professionals, teachers, tour guides, pilots and air traffic controllers); placement tests for students enrolling for language study (Arabic, Chinese, French, German, Indonesian, Italian, Japanese, Russian

and Spanish) within the university context; national government-funded projects investigating issues such as the impact of language background and programme exposure on student achievement in Asian languages in Australian schools; and the language proficiency requirements of interpreters and translators applying for national certification. The centre also undertook major contract research for larger international testing agencies such as TOEFL, IELTS and Pearson in relation to their large-scale, internationally recognised commercial English tests.

In addition to building assessment tools and undertaking research, the LTRC contributed significantly to building language assessment literacy via publications such as the *Dictionary of Language Testing* (Davies et al 1999) and the *Mark My Words* video series (Davies et al 1996)[5], as well as through the *Melbourne Papers in Language Testing* – a major publishing outlet for research in language testing from 1992, including for PhD students. More recently, the centre has been at the forefront of language assessment capacity-building in Australia and the wider region through its role in the 2012 establishment of the Association for Language Testing and Assessment of Australia and New Zealand (ALTAANZ), a regional organisation to promote best practice in language testing and assessment in educational and professional settings. The *Melbourne Papers* have since been adopted as the association's peer-reviewed journal, *Papers in Language Testing and Assessment*. The centre recently launched an online Professional Certificate in Language Assessment designed for teachers and practitioners interested in learning more about language testing and assessment.

Through its research and test development activity, the centre helped to shape theory and practice on a wide variety of issues, both within Australia and beyond: performance testing and Rasch analysis (McNamara 1996, McNamara, Knoch and Fan 2019); test validity and validation (Davies and Elder 2005, Knoch and Chapelle 2018); test ethics, test fairness and social consequences (Davies 1997, Elder 1997, McNamara and Ryan 2011); rater and interlocutor behaviour (Brown 2003, Brown and Hill 1998, Knoch 2011, Lumley 2002); language assessment in higher education contexts (Elder and Read 2015, Knoch and Elder 2013, O'Hagan 2014); assessing speaking proficiency (Iwashita, Brown, McNamara and O'Hagan 2008, McNamara 1997, O'Loughlin 2001); classroom-based assessment (Hill and McNamara 2012, Knoch and Macqueen 2017); the testing of language/s for specific purposes (Elder 2001, 2016, Knoch and Macqueen 2019, Pill 2016); and policy formation and the social dimensions of language assessment more generally (Frost and McNamara 2018, McNamara and Roever 2006, Roever and Wigglesworth (Eds) 2019). Many of the authors cited above also completed

5 This video series is now available to view via the *Resources* section of the ILTA website: www.iltaonline.com

their PhDs at the University of Melbourne, producing an impressive list of home-grown LTA researchers, including Tim McNamara, Cathie Elder, Kieran O'Loughlin, Tom Lumley, Annie Brown, Noriko Iwashita, Lyn May, Luke Harding, Kathryn Hill, Sally O'Hagan, Kellie Frost, Susy Macqueen and John Pill, among others.

Centre for Research in Testing, Evaluation and Curriculum (CRTEC), University of Roehampton, Surrey

The Centre for Research in Testing, Evaluation and Curriculum (CRTEC) was established by Cyril Weir in the early 2000s following his move from the University of Reading to take up his first professorship at the University of Surrey, Roehampton (later the University of Roehampton). Barry O'Sullivan joined him there after being appointed as Reader. In 2005, the name of the centre was changed to the Centre for Language Assessment Research (CLARe) and it continued to operate until around 2014 by which time both Cyril and Barry had moved on to work in other institutions – Cyril as a professor at the University of Bedfordshire (see the section 'Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire') and Barry at the British Council, where he later became Head of Assessment Research and Development for English and Exams. While at Roehampton, Cyril Weir also took on the role of joint Series Editor (with Mike Milanovic) for the Cambridge English Studies in Language Testing (SiLT) series, helping over 35 academic volumes to reach publication – on all aspects of LTA and including quality PhDs.

During its approximately 12-year life, CRTEC/CLARe offered PhDs in Language Testing and co-supervised on PhDs in other areas with colleagues from the Education Faculty. PhD graduates included Jessica Wu (now Programme Director of the R&D Office at the Language Training and Testing Center (LTTC) in Taiwan), Akmar Saidatul Abidin (subsequently founder of the Malaysian Association for Language Testing) and Elif Kantarcıoğlu (Dean of Bilkent University School of English Language, Turkey). Short training courses in language testing, similar to those run at Reading in the 1990s, were also offered during the summer months and these became a model for some of the training courses offered by Cambridge English and the Association of Language Testers in Europe (ALTE) from about 2005 onwards.

Though quite small in terms of number of staff and students, Barry O'Sullivan recalls CRTEC/CLARe being involved in a wide range of practical test design and development projects, providing consultancy and research support to various ministries, universities and test development agencies around the world (personal communication, 2019). Clients included:

UK-based testing organisations, such as Cambridge English, British Council, Trinity College and City and Guilds; and international agencies, such as Veracruz University in Mexico, Bilkent University in Turkey, Sultan Qaboos University in Oman, Zayed University in UAE and the Ministry of Education in Egypt. PhD students at the centre were routinely involved in project and consultancy work, not just in testing English but also in assessing Art and Science at secondary level (in Portugal and Sri Lanka, respectively). A series of PhD seminars provided a chance for students to not only present on their own research, but also participate in discussions about the emerging concept of a socio-cognitive approach in language assessment.

Barry recalls that, while he was doing much of the travelling to support the various research and consultancy projects commissioned from CRTEC, Cyril was able to devote considerable time and space to develop his thinking and writing on the socio-cognitive framework approach, work they originally started while both at Reading. The approach sought to assemble and articulate differing dimensions of validity (context, theory-based, scoring, criterion-related and consequential) all of which need to be considered and addressed in any test development and validation enterprise. CRTEC's research and consultancy projects offered a rich opportunity to apply and trial the socio-cognitive approach, shaping further thinking and helping to refine the models for speaking, reading, writing and listening. Barry believes it also brought the approach to a global audience through the projects and through their dissemination activities. This eventually led to the publication of Cyril's 2005 volume *Language Testing and Validation: An Evidence-based Approach* just as he was about to move on to his next post. Barry comments as follows on their time at CRTEC (personal communication, 2019): 'When we were in the same space, we had ample time to discuss the frameworks and found plenty of time to agree (and disagree) on the structure and details – I'm not sure the socio-cognitive approach would have become so influential without that time together and without the excellent PhD students we had there.' A fuller discussion of the origins and evolution of the socio-cognitive framework in the early 2000s in relation to speaking assessment can be found in Barry O'Sullivan's chapter in this volume (Chapter 2).

The Canadian 'academic network' of university-based institutions

Carolyn Turner recalls that, when she entered the world of language testing and assessment in 1984 to undertake a PhD at McGill University, the wealth and richness of the field in Canada was found more through individuals at academic institutions and collegial work across institutions than through specific academic LTA units or centres (personal communication, 2019). To this day, she notes that there continue to be few academic institutions that

have a unit or centre dedicated to language testing, though one example might be the Language Assessment Sector within the Official Languages and Bilingualism Institute (OLBI) at the University of Ottawa. Measurement and evaluation university departments and units do exist across Canada, but typically have a more general focus (e.g. Centre for Research in Applied Measurement & Evaluation at the University of Alberta). For this reason, language testing research tends not to be centralised within academic institutions but is instead promoted through individual university professors and staff via their university teaching, research, graduate student training, publications, through consultancy with provincial ministries of education concerning student assessment and provincial exams, and through consultancy with the Canadian government on bilingual assessment issues since Canada has two official languages – English and French.

The Canadian government policy on bilingualism includes language tests at the civil servant level, in immigration/citizenship, etc., but each one of these areas functions separately and sometimes draws on university experts to participate on committees focusing on policy and/or test development. There is also co-operation between government and academia on standards specifically for language competence and test development. Perhaps the best-known example is the *Canadian Language Benchmarks/Niveaux de compétence linguistique canadiens* project led by Bonnie Norton (University of British Columbia) and Michel Laurier (Université de Montréal). They were originally developed in the mid-1990s as a common framework for the description and evaluation of the language proficiency of adult newcomers to Canada. Through evolving research and practice, they presently serve as national standards for French and English language proficiency in educational, training, community and workplace settings (e.g. healthcare), and also are cited and consulted internationally.

University-based language testers across Canada were also instrumental in developing standardised language proficiency tests, some of which are well recognised and used internationally. For example, the Canadian Academic English Language (CAEL) Assessment, originally developed and managed by Janna Fox at Carleton University, was designed as an alternative to traditional, discrete-point, multiple-choice proficiency tests in verifying whether a test taker's level of English was adequate to meet the demands of academic study at college or university level. A criterion-referenced, topic-based performance test, the CAEL was the first high-stakes proficiency test to utilise fully integrated tasks, which maintained a central topic throughout the paper-based reading, listening and writing sections; the speaking section of the test, known as the CAEL Oral Language Test (OLT), was task based and computer administered. The new version of the test, CAEL CE (Computer Edition) was launched in 2017 by Paragon Testing Enterprises, a subsidiary of the University of British Columbia (UBC), where the test is now housed

and administered. In 2018, the original paper-based version of CAEL was retired. Paragon Testing Enterprises is also responsible for the Canadian English Language Proficiency Index Program (CELPIP) Test which provides proof of English language proficiency for immigration to Canada or for Canadian citizenship. Both the CAEL CE and CELPIP are administered across Canada and internationally. CanTEST/TESTCan, available in both English and French, was developed as a test forming part of the admission procedures to Canadian postsecondary institutions or for professional licensing purposes. Canadian language testers working on CanTEST/TESTCan included Margaret Des Brisay, Doreen Bayliss and Mari Wesche, all based at the Second Language Institute, University of Ottawa. In 2007 the institute became OLBI, with its own Language Assessment Sector, which continues to administer the test under its current Director, Beverly Baker.

Given the strong focus on language (including bilingual) pedagogy and assessment in the Canadian academic context, two other important areas of research and scholarly interest relate to washback and classroom-based assessment (CBA). Building on the work of Alderson and Wall at Lancaster, Liying Cheng (now professor at Queen's University, Kingston, Ontario) focused attention on washback in a wide range of educational contexts, stimulating research studies in this area at the classroom level which were reported in conference symposia and publications (Cheng 2005). For many years, research concerning CBA, classroom teachers, student and teacher feedback and CBA's relation to external tests was relatively sparse in the LTA literature. However, from the 1990s onward a burst of CBA-related research came out of Canada including reference books for educators (e.g. Genesee and Upshur 1996, Cheng and Fox 2017). Within this research, innovative methodologies were added to the previously dominant quantitative methods; paradigms such as qualitative and mixed methods research became evident and appropriate means to investigate language assessment issues. Key researchers in this regard included several individuals from the Ontario Institute for Studies in Education (OISE) at the University of Toronto (Merrill Swain, Alister Cumming and Eunice Jang) as well as, among others, Liying Cheng (at Queen's), Janna Fox (at Carleton), Michel Laurier (at Université de Montréal and Ottawa), Carolyn Turner (at McGill) and John Upshur (at Concordia University in Montreal).

Growing interest in qualitative and mixed research methods in Canadian language testing research was reflected in pre-conference workshops at the academic language testing events hosted by Canadian language testers, and was also complemented by a strengthening of expertise in applying quantitative methodologies, resulting from interdisciplinary work with colleagues in the fields of statistics, psychometrics, validity theory and studies of the mathematical basis of classical test theory and measurement error models. Bruno Zumbo, from the University of British Columbia, remains a key figure in this area.

The academic personnel and their university affiliations mentioned above mean that LTA-related courses have been promoted, taught and in many cases developed in academic units mainly due to the presence of these academics. As a result, a new generation of language testers emerged in Canada, many of whom are now making their own contributions to the field in new areas of research and scholarship. They include, among others, Beverley Baker (with a focus on language assessment literacy), Khaled Barkaoui and Heike Neumann (with a focus on writing assessment), Christine Doe (with a focus on diagnostic assessment) and Talia Isaacs (with a focus on pronunciation assessment).

Between 1983 and 2015 Canada successfully hosted five LTRCs at which Canadian language research was highlighted (see Fox et al (Eds) 2007) and the strength of the Canadian network is reflected in the founding in 2009 of CALA/ACEL (Canadian Association of Language Assessment/*Association canadienne pour l'évaluation des langues*). Canada's influence on and contribution to LTA theory and practice is further reflected in its international collaboration with testing organisations, such as ETS and Cambridge English (the developers of TOEFL and IELTS), and with agencies such as the International Civil Aviation Organization (ICAO), which has its main headquarters in Montreal.

Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire

CRELLA came into being in 2005 when Cyril Weir first arrived at the University of Luton, although the university was renamed the University of Bedfordshire in August 2006. The centre expanded with the addition of Roger Hawkey (fellow student and colleague of Cyril's in the 1980s) and Anthony Green, now Director of CRELLA and a former PhD student of Cyril's at CRTEC before joining Cambridge English. By the time of the UK Research Assessment Exercise in 2008, CRELLA had announced itself as one of the leading centres of research excellence within the university.

The work of the centre was originally built around the socio-cognitive framework first elaborated in Weir (2005). CRELLA set out on a long-term research programme to apply and further develop the socio-cognitive framework in collaboration with test providers around the world. The framework was used as a basis for the validation of internationally recognised English tests that covered different proficiency levels, skills and domains. Organisations that CRELLA has worked with in the UK include, among many others: the British Council, Cambridge English and Trinity College London. Internationally, they have included the LTTC in Taiwan, the Society

for Testing English Proficiency (STEP) in Japan, and a wide range of universities and other test providers in Europe, Asia and South America.

Cyril took with him to CRELLA the strong connection with Cambridge English that he had maintained since his years at Reading. The centre was a founding partner of English Profile, a long-term programme of collaborative research with Cambridge English to provide the Council of Europe with Reference Level Descriptions that complement the description of proficiency levels in the Common European Framework of Reference for Languages (CEFR), relating these specifically to English language education and assessment. By 2014, CRELLA had expanded to include nine researchers including Stephen Bax, John Field, Liz Hamp-Lyons, Fumiyo Nakatsuhara and Lynda Taylor. The centre had taken on a broader remit to investigate assessment literacy, learning-oriented language assessment and academic language proficiency, as well as elements of the socio-cognitive framework.

Soon after arriving at Luton, Cyril established doctoral research opportunities at CRELLA. Nick Saville recalls transferring his PhD with Cyril from Roehampton to Luton and comments that 'in CRELLA the academic milieu became a shining example within the Luton/Bedfordshire set up – again Cyril was innovative and led the way' (personal communication, 2019). CRELLA PhD graduates include Nick Saville (Director of Research and Thought Leadership at Cambridge English), Rachel Yi-fen Wu (LTTC, Taiwan), Sathena Chan (now Senior Lecturer at the university), Jamie Dunlea (Senior Researcher at the British Council), Daniel Waller (Senior Lecturer in ELT, Testing and TESOL at the University of Central Lancashire) and Mark Chapman (Director of Test Development at the WIDA Consortium).

Despite the sad loss of both Stephen Bax and Cyril Weir in 2017 and 2018 respectively, the centre continues to thrive. Mike Milanovic joined the CRELLA team as Visiting Professor in 2014 following his retirement from Cambridge English. Recent work at CRELLA includes a systematic review of the four IELTS academic modules for the IELTS partners, a new diagnostic test of academic reading and writing skills for all incoming students at Bedfordshire (see the chapter by Chan and Latimer in this volume) and a series of investigations into the impact of technology on tests of spoken interaction.

The impact and legacy of academic institutions

Reflecting on the brief pen portraits of the academic institutions and contexts given above, there can be little doubt regarding the significant and enduring role they have played in the field of LTA in many different ways. This final section of the chapter highlights a number of common themes that emerge, drawing together some threads to consider impact and legacy under eight main headings.

1. The evolution of language testing and assessment theory

As we might hope and expect from any academic institution, the LTA research and consultancy activity conducted in/between university-based centres and departments has helped to advance our theoretical knowledge and understanding in some key areas.

In some cases, existing theory and knowledge were significantly expanded: for example, both Edinburgh and UCLA created firmly grounded links between the fields of linguistics, applied linguistics and LTA; they also highlighted the importance of using both qualitative and quantitative methodologies in LTA research, e.g. techniques such as discourse analysis and item response theory (IRT). In other cases, institutions, or individuals within/across institutions, sometimes led the way in introducing major new directions for the field. Researchers at Reading pioneered the European interpretation of communicative language testing, while Lyle Bachman drew on Canale and Swain's Model of Communicative Competence for his test methods framework. Lancaster led the way for an exploration of test washback and impact, and later developed approaches to diagnostic assessment. Alan Davies and academic colleagues in Australia (as well as in Israel) advanced our understanding of ethics in testing and the social dimensions of assessment. Cyril Weir and colleagues at CRTEC and CRELLA presented a unified socio-cognitive approach to test development and validation. Without the work of individuals and teams in specific university-based contexts around the world it is questionable whether LTA thinking and theory would have advanced as much as it has over the past half century.

2. The development of a professional research network and community

Individual applied linguists with an interest in LTA often found themselves working in ones or twos within a single academic institution. They were therefore strongly motivated to find ways of meeting together professionally to share their ideas and challenges. It is notable that the very first meeting of the Language Testing Research Colloquium (LTRC) in February 1979 was held fairly informally in a Boston hotel room, attended by (mostly) academics from a variety of different institutions, mainly, but not exclusively, in the US. As Dan Douglas recalls: 'They were indeed a select group of applied linguists who shared an academic interest in the theory and practice of second language assessment' (2015:6). Participants came together from their various academic contexts, primarily to present papers on their own testing research and to engage in discussion of topics and problems they shared in common, including: the nature of communicative competence, the effectiveness of research methodologies and the impact of technological developments.

As Lyle Bachman and Adrian Palmer reported when relating the story of LTRC, it was 'both surprising and gratifying to find that there were others who shared their interest in this abstruse and relatively unpopular enterprise' (Douglas 2015:6).

Just over a decade later, in 1992, the International Language Testing Association (ILTA) was 'born' at LTRC in Vancouver, Canada, again with significant involvement from academics across the worldwide professional community, and from 1997, LTRC was adopted as the ILTA research conference. Although for the first 14 years annual LTRC meetings were held exclusively on the North American continent (with the exception of Honolulu in 1982), from 1993 onwards the Colloquium found venues in all parts of the world, often hosted by a specific university and organised by a local team of faculties and students, with some support from the ILTA Board. In 1993, LTRC was located in Europe for the first time (Cambridge and Arnhem), followed by Asia in 1999, Australasia in 2006, and South America in 2018. Outreach into new regions to extend the professional research network and to offer opportunities to an ever-widening community remains a key aspiration of ILTA. At the time of writing, LTRC 2020 is to be hosted for the first time on the African continent, in Hammamet, Tunisia – another 'first' for the field.

LTRC is not the only regular gathering of language testers, however. British language testers from across different academic UK-based institutions first assembled for a networking weekend in Lancaster in 1980 – a year after the first LTRC took place in the US. The UK's Language Testing Forum (LTF) subsequently became a fairly informal annual event and is now the annual conference of the recently formed UK Association for Language Testing and Assessment (UKALTA), hosted each year by a British university.

Over the years it has often been university-based language testing researchers and scholars who have initiated national or regional associations to facilitate the sharing of research findings and assessment practice. In the 1980s and early 1990s there was relatively little engagement from the UK language testing community in European matters, with the notable exception perhaps of links between Mike Milanovic (and the team at Cambridge English) and fellow language test providers based within key European universities and testing agencies. These links led in 1989 to the establishment of the Association of Language Testers in Europe (ALTE) by the universities of Cambridge (UK) and Salamanca (Spain), who shared a common interest in designing and delivering language assessments (tests of English and Spanish, respectively).

In the years that followed, other European testing providers joined the ALTE organisation to share in discussions and to collaborate on projects that could help improve their own language assessment theory and practice. Two specific projects that benefitted the wider LTA community were development of the ALTE Code of Practice and Principles of Good Practice, first available

in 1994; these were followed by the publication of a *Multilingual Glossary of Language Testing Terms* in 1998. Training courses and events for ALTE members date back to the 1990s – often in association with Cyril while he was still at Reading University; his direct involvement in ALTE conferences and other events grew during his time at Roehampton and CRELLA. Cyril was always a staunch supporter of ALTE's initiatives in developing assessment literacy in multilingual contexts and he regularly taught on the ALTE training courses held throughout Europe between 2005 and 2016.

In 2004, Charles Alderson and colleagues at Lancaster played a significant role, together with fellow European language testers from Finland, the Netherlands, Spain, Greece and Germany, in the creation of the European Association for Language Testing and Assessment (EALTA). More recent examples of regional LTA associations include the foundation of CALA/ACEL in 2009, the Association for Language Testing and Assessment of Australia and New Zealand (ALTAANZ) in 2012 and the Asian Association for Language Assessment (AALA) in 2014, as well as UKALTA in 2016.

The past 50 years, therefore, have seen the evolution of a global LTA professional community as well as regional networks, many of whom hold annual conferences and other occasional events. ALTE is a now regional association of 34 organisational members, all language test providers representing 25 European languages, as well as 58 institutional and 500 individual affiliates; it holds annual conferences and training courses. In a similar way, EALTA offers regular summer schools and pre-conference workshops, and for some years now ILTA has sponsored regional workshops as part of its awards programme, often in new, relatively 'unreached' parts of the world. Such networks, often initiated and facilitated by academic language testers, enable them to find colleagues interested in the same types of issues, leading to collaboration through joint research projects, publications, presentations, training workshops, etc. As Carolyn Turner commented from the Canadian perspective (personal communication, 2019): 'The international involvement influences and feeds on the local involvement and vice versa.'

3. Collaborative research endeavours

The early, often more informal, networking of individual language testers across academic institutions resulted in collaborative approaches to LTA research and practical test development activity which have endured to the present day, sometimes leading to major research programmes and test development outcomes.

In the US in the late 1950s and early 1960s, for example, Robert Lado and his contemporaries collaborated on the development of Test of English as a Foreign Language (TOEFL) for ETS. Over the following decades, successive generations of university-based academics served on the TOEFL Committee

of Examiners, undertook commissioned research into the test, much of which was published for a wider audience, and contributed to the redevelopment of the test in the TOEFL 2000 and Internet-based Test of English as a Foreign Language (TOEFL iBT) projects some 40 years later.

In a similar way, the 1970s and 1980s saw academics from Edinburgh, Lancaster and Reading universities collaborating on the development and validation of the English Language Testing Service (ELTS). Building on this work, towards the end of the 1980s the International English Language Testing System (IELTS) was developed with academic input from Australia (David Ingram and Elaine Wylie) as well as the UK; and when IELTS was revised in 1995 it was with the professional advice and support of academics in Australia, New Zealand and the US (David Ingram, John Read and Lyle Bachman). A similar story can be told regarding the design and development of the Occupational English Test (OET) in Australia in the late 1980s and 1990s, in which Tim McNamara and Australian colleagues played a significant role.

From the 1970s to the 1990s, opportunities for academics to engage in substantive LTA research tended to be restricted to research projects specially commissioned and funded by large commercial test providers such as ETS and Cambridge English. Examples of such projects include those mentioned above to develop TOEFL and ELTS/IELTS. However, as the professional research network and community expanded and as test providers took more seriously their responsibility to bring forward test validation evidence for the benefit of test stakeholders, there was a marked growth in grant-funded programmes administered by the commercial testing agencies. ETS, Cambridge English, British Council and Pearson all now routinely offer annual funding opportunities that enable university-based LTA academics, especially early career researchers, to undertake important test validation research, often in collaboration with colleagues in their own or other university contexts. Testing programmes around the world that stand out for their commitment to quality are those that have involved LTA academics. The recently published volume *English Language Proficiency Testing in Asia: A New Paradigm Bridging Global and Local Contexts* by Su, Weir and Wu (Eds) (2019) testifies to the extent to which Cyril Weir and other LTA academics have played a key role in encouraging and supporting research-based test development and validation expertise in other regions of the world.

4. The growth of a postgraduate research cadre

From the 1960s to the 1990s, academic institutions in the US, the UK, Canada and Australia were the predominant locations for accessing a Master's or doctoral level programme of study in LTA. In the 1970s and 1980s, Edinburgh, Reading and UCLA featured high up the list in this regard, offering the

means for the next generation of language testers to be trained and gain postgraduate research experience. Over the next 30 years, however, Lancaster, Melbourne, Roehampton, CRELLA and some of the Canadian universities developed a wide range of onsite or distance-based, full- or part-time courses in LTA, as well as providing regular summer schools, seminars and workshops for postgraduate students.

Offering academic tuition and practical training and experience to international students in more flexible ways thanks to new educational technologies means that that LTA expertise can now be accessed and supported more widely than ever before around the world, especially across the continents of Europe and Asia. The current CRELLA webpage illustrates the extent of reach nowadays: for the period 2009–19 the centre lists its successful PhD students as coming from: UK, Germany, Holland, Poland, Georgia, Jordan, Egypt, Turkey, Pakistan, Sri Lanka, Malaysia, Thailand, China, Japan, Hong Kong, Taiwan and Australia. In many cases, a CRELLA PhD study investigated some aspect of the student's in-country pedagogical or assessment context, ensuring that any research findings would have direct relevance to their home situation and that the research training they acquired might benefit their local educational environment.

5. The dissemination of LTA knowledge and practice

Throughout the 1960s and even into the 1970s, the number of LTA-related textbooks available for language testers or language teachers to use when designing or interpreting a test was relatively limited. Those that did exist were sometimes quite specialised or technical. From the late 1980s onwards, however, as language testing and assessment systems began to play a greater role in higher, secondary and even primary education, both nationally and internationally, accessible material was needed to promote a better understanding of the principles and practice of testing across a larger constituency within education and society, particularly among language teachers and test writers. It was predominantly university-based academics who took it upon themselves to try and communicate the theory and practice of LTA through publications that could be accessed by a more general audience. Early examples include books by Hughes (1989), Davies (1990), Weir (1990, 1993), Bachman (1990), Alderson et al (1995), McNamara (1996) and Bachman and Palmer (1996).

This first wave of titles on general testing theory and practice was quickly followed by a series of more focused volumes in the later 1990s and 2000s addressing the testing of specific language skills such as reading or writing, grammar or vocabulary (e.g. Alderson 2000, Read 2000, Weigle 2002, Purpura 2004), as well as volumes focusing on emerging areas such as assessing language for specific purposes, the testing of young language learners and

the application of technology in assessment (Douglas 2000, McKay 2005, Douglas and Chapelle 2006). Once again, it was typically university-based academics who were able to combine the knowledge and experience needed with the writing/editing time and skills required to produce such publications.

The growing use of language tests at all levels of education and in a range of social contexts (e.g. immigration and citizenship, licensing of internationally qualified health professionals) cast a spotlight on issues concerning the use, misuse and potential abuse of tests. Building partly on the 1990s research into test washback and impact, including a growing recognition of the multiple stakeholder communities involved in assessment, academic language testers in Israel and Australia addressed the power dynamics of language testing in society (Shohamy 2001, McNamara and Roever 2006), while others explored what it might mean to behave ethically as language testers. Together with a team of ILTA colleagues, Alan Davies, already well known for his interest in ethics (Davies 1997), worked to develop a code of ethics (2000, updated 2018) for the LTA field, followed by a set of guidelines for good practice, the ILTA *Guidelines for Practice in English* (2007). The ILTA guidelines acknowledge earlier work by academics in the Japanese Language Testing Association (JLTA) to draft a similar protocol for their context, and in the European arena EALTA produced its own *Guidelines for Good Practice in Language Testing and Assessment* in 2006. Both the ILTA code and the EALTA guidelines have been translated into over 30 languages so that a wide range of test developers and users can access the core principles and practices that characterise good-quality assessment.

International and regional associations, such as ALTE, ILTA and EALTA, continue the important work of LTA outreach, often involving staff from academic institutions in the design of materials, the delivery of presentations and the running of training courses (cf. ALTE Training Courses, ILTA Workshops and EALTA Summer Schools) as part of the process of developing at the local, regional and national level what has come to be known as 'assessment literacy' (Taylor 2009, 2013).

6. The dissemination of LTA publications in promoting assessment literacy

The promotion of assessment literacy among an ever-widening constituency of language testing stakeholders has been facilitated by the expanding volume of published material on language testing and assessment available in the public domain, much of it in printed form but increasingly in other media.

Books (including academic monographs, edited collections, technical handbooks and testing encyclopaedia) and research reports (both hard copy and online) have been widely published since half a century ago when only a

handful of titles were available. The appearance over recent years of edited encyclopaedia or handbook volumes testifies to how far the field of LTA has broadened and diversified to embrace multiple areas of specialism. Such volumes are typically edited collections of chapters by LTA academics from all over the world; examples include Shohamy and Hornberger (Eds) (2008), Coombes, O'Sullivan, Stoynoff and Davison (Eds) (2012) and Tsagari and Banerjee (Eds) (2016).

In 1984, Arthur Hughes and Don Porter founded *Language Testing*, the first academic journal dedicated to language testing and assessment, and almost 20 years later it was joined by a sister journal, *Language Assessment Quarterly,* founded by Antony Kunnan in 2003. Between them these two journals have published hundreds of peer-reviewed papers in the field. Other organisations or regional networks publish collections of working papers (e.g. LTRC Melbourne) and similar research outputs.

The 1990s saw the publication of the first *Multilingual Glossary of Language Testing Terms* (produced in 1998 by ALTE members, many of whom were working in academic contexts across Europe) and the *Dictionary of Language Testing* (produced in 1999) by Alan Davies and his colleagues at LTRC, Melbourne (both were published as part of the SiLT series). The academic team at LTRC Melbourne was also responsible for creating an innovative 6-part video series on assessing second and foreign language skills entitled *Mark My Words* (Davies et al 1996). A similar project was a series of videos of ILTA members giving short lectures on various aspects of language testing practice and theory. The project was headed up by Glenn Fulcher and Randy Thrasher in 1999, funded by ILTA and International Christian University in Japan. Many of the videos were made during the LTRC 1999 meeting in Tsukuba, Japan.

ILTA was also responsible for the ILTA Language Testing Bibliography project, begun in 1999 by Caroline Clapham and Dianne Wall, an invaluable resource for the language testing community and one that has lasted to the present day. With entries ordered according to topic and author, the ILTA Bibliography was originally published only in print form, but was later made accessible to ILTA members via the organisation's website. An additional bibliography of doctoral dissertations on language testing was compiled and published in 2008; it is regularly updated and is also available from the ILTA website.

Initiatives such as the above help to support not only the professionalisation of those in the immediate language testing and assessment community, but also the growth of assessment literacy among a wider circle of test stakeholders, such as university admissions tutors, national policy-makers and regulatory authorities.

7. Practical language test development and validation activity

Language testers or testing centres within an academic institution are often tasked by their university with creating or managing a test that will meet the specific needs of their local context, e.g. a selection test for university admission, or a test for certifying international teaching assistants. Examples over the past 50 years include: EPTB (at Birmingham) and ELBA (at Edinburgh) in the 1960s; ELTS/IELTS (at Edinburgh/Lancaster), TEAP/TEEP (at Reading), the Taped Evaluation of Assistants' Classroom Handling (TEACH) (at Iowa State) and CAEL (at Carleton, Ottawa) in the 1970s/1980s; and, most recently, Bedfordshire Academic Reading Test (BART) (CRELLA) in the 2000s (see the chapter in this volume by Chan and Latimer). Test development projects of this nature can provide valuable opportunities for doctoral students to gain applied knowledge and experience of the process of test design and validation.

As well as developing larger-scale tests such as these, some of which gain wider currency beyond the originating university, academic language testers are often invited to advise other faculty staff on internal assessment frameworks for their students or to work on developing a range of test tools specific to a disciplinary area. Sometimes their expertise is sought by university admissions staff on the cut-scores required for entry to a university course, though anecdotal evidence suggests this perhaps does not happen as often as it should and efforts continue to engage more effectively with both internal and external test-stakeholders. Ensuring the proper interpretation and use of test scores by an academic institution or any other organisational context (e.g. department of immigration, medical licensing regulator) remains an ongoing challenge.

Beyond their own institutional context, university-based language testers are often engaged as consultants to advise organisations or governments on their LTA needs, whether as part of a national educational reform programme or as part of new regulatory procedures in areas such as public health, civil aviation or migration/citizenship. Lancaster University, LTRC (Melbourne), CRTEC/CLARe (Roehampton) and CRELLA (Bedfordshire) have all had significant involvement in some or all of these areas over the past 30 years.

8. Advocacy and expertise in relation to social policymaking and implementation

Consultancy projects such as those referred to above testify to the important role that language testers sometimes play in wider society, with the potential to impact directly on educational and social policy, decision-making and implementation.

The team at LTRC in Melbourne, for example, were instrumental in helping

to develop, implement and validate the OET for evaluating the English language skills needed by doctors, nurses and other health professionals working in the Australian context (McNamara 1996). In 2014–15 the CRELLA team undertook a research project with the UK General Medical Council (GMC) to provide an evidence-based analysis of English language proficiency tests for potential use within their medical registration and licensing procedures (Taylor and Chan 2015). Similar health-related consultancy has been undertaken by academics in Canada using the Canadian Language Benchmarks as a reference point in Healthcare Access for Language Minorities research (HCALM) funded by Health Canada. The research promoted enhanced language ability for health practitioners working with minorities and led to international co-operation in the health communication sciences, resulting in conference papers and symposiums at LTRC in 2010 and 2014 as well as relevant publications (e.g. Isaacs, Laurier, Turner and Segalowitz 2011).

As a professional organisation, ILTA members from the academic community played a major role in advising the international civil aviation authorities on English language standards for pilots and air crew, and other consultancies have been carried out with the military.

Conclusion

Our aim in this chapter has been to examine and reflect upon the impact and legacy of academic institutions within the field of language testing and assessment. We have focused on just a handful of institutions to consider their role in developing LTA theory and practice and in shaping an understanding of the place and purpose of language assessment within education and society more broadly. As we explained at the outset, for illustrative purposes we limited our scope to specific contexts in the UK, US, Canada and Australia, and our coverage can therefore only be partial and selective. We fully recognise that there are numerous other academic institutions, teams and individuals around the world who could be considered and added to the list, not to mention many lone researchers, as well as other providers of tests of English and other languages.

Cyril Weir spent the bulk of his professional and academic life as a language tester committed to working within various university-based contexts in the confident belief that such contexts, through research, scholarship and consultancy, could make a significant contribution to LTA knowledge and expertise with positive impact for education and wider society. Cyril came from an exam board and EFL teaching background and he soon recognised the need for theory-informed practice and practice-informed theory. Though he valued independence from the big testing organisations, Cyril was always ready to engage proactively with assessment agencies as well as with individual enthusiasts because he believed in the benefits of team working and that

bringing together specialists in different aspects of language testing research and practice can achieve positive outcomes for all. The story we have told here suggests that Cyril was correct in his assessment. *Gratias maximas tibi agimus, Cyrillus – honoramus te*!

Acknowledgements

We are most grateful for the help of a number of language testing colleagues around the world who kindly provided information and input in the form of official historical accounts of their centres/departments combined with personal recollections and reflections, and who checked an early draft for accuracy and balance. Particular thanks must go to: Tineke Brunfaut, Sara Cushing, Cathie Elder, Janna Fox, Luke Harding, Ute Knoch, Barry O'Sullivan, Jim Purpura, Nick Saville, John Slaght, Carolyn Turner, Dianne Wall and Eddie Williams.

References

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

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

Alderson, J C and Clapham, C (Eds) (1992) *IELTS Research Report 2: Examining the ELTS Test – An Account of the First Stage of the ELTS Revision Project*, Cambridge: British Council/UCLES/IDP.

Alderson, J C, Clapham, C and Wall, D (1995) *Language Test Construction and Evaluation*, Cambridge: Cambridge University Press.

Allen, J P B and Davies, A (Eds) (1977) *Testing and Experimental Methods*, Edinburgh Course in Applied Linguistics volume 4, Oxford: Oxford University Press.

Allen, J P B and Pit Corder, S (Eds) (1974) *Techniques in Applied Linguistics,* Edinburgh Course in Applied Linguistics volume 3, Oxford: Oxford University Press.

Association of Language Testers in Europe (ALTE) (1998) *Multilingual Glossary of Language Testing Terms*, Studies in Language Testing volume 6, Cambridge: UCLES/Cambridge University Press.

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

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

Brown, A (2003) Interviewer variation and the co-construction of speaking proficiency, *Language Testing* 20 (1), 1–25.

Brown, A and Hill, K (1998) Interviewer style and candidate performance in the IELTS oral interview, in Woods, S (Ed) *Research Reports 1997 Volume 1,* Sydney: ELICOS, 173–191.

Brown, J D, Hudson, T, Norris, J M and Bonk, W (2002) *An Investigation of Second Language Task-based Performance Assessments,* Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Brumfit, C (2001) Alan Davies and British applied linguistics, in Elder, C, Brown,

A, Grove, E, Hill, K, Iwashita, N, Lumley, T, McNamara, T and O'Loughlin, K (Eds) *Experimenting with Uncertainty: Essays in Honour of Alan Davies*, Studies in Language Testing volume 11, Cambridge: UCLES/Cambridge University Press, 2–4.

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

Cheng, L and Fox, J (2017) *Assessment in the Language Classroom: Teachers Supporting Student Learning*, London: Palgrave.

Clapham, C (1996) *The Development of IELTS: A Study of the Effect of Background Knowledge on Reading Comprehension*, Studies in Language Testing volume 4, Cambridge: UCLES/Cambridge University Press.

Clapham, C and Alderson, J C (Eds) (1997) *IELTS Research Report 3: Constructing and Trialling the IELTS Test,* Cambridge: The British Council/UCLES/IDP.

Coombes, C, O'Sullivan, B, Stoynoff, S and Davison, P (Eds) (2012) *Cambridge Guide to Second Language Assessment*, Cambridge: Cambridge University Press.

Davies, A (1968) *Language Testing Symposium*, Oxford: Oxford University Press.

Davies, A (1990) *Principles of Language Testing*, Oxford: Basil Blackwell.

Davies, A (1997) The limits of ethics in language testing, *Language Testing* 14 (3), 235–241.

Davies, A (2008) *Assessing Academic English: Testing English Proficiency 1950–1989 – The IELTS Solution*, Studies in Language Testing volume 23, Cambridge: UCLES/Cambridge University Press.

Davies, A and Elder, C (2005) Validity and validation in language testing, in Hinkel, E (Ed) *Handbook of Research in Second Language Teaching and Learning*, Mahwah: Lawrence Erlbaum, 795–813.

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

Davies, A, Brown, A, Elder, C, Evans, R, Grove, E, Iwashita, N and McNamara, T (1996) *Mark My Words: Assessing Second and Foreign Language Skills. 6-part Video Series*, Melbourne: Multimedia Production Unit, The University of Melbourne.

Douglas, D (2000) *Assessing Language for Specific Purposes*, Cambridge: Cambridge University Press.

Douglas, D (2015) *The Language Testing Research Colloquium and the International Language Testing Association: Beginnings*, available online: https://cdn.ymaws.com/www.iltaonline.com/resource/resmgr/docs/A_Short_History_of_LTRC.pdf.

Douglas, D and Chapelle, C A (2006) *Assessing Language through Computer Technology*, Cambridge: Cambridge University Press.

Elder, C (1997) What does test bias have to do with fairness?, *Language Testing* 14 (3), 261–277.

Elder, C (2001) Assessing the language proficiency of teachers: Are there any border controls?, *Language Testing* 18 (2), 149–170.

Elder, C (2016) Exploring the limits of authenticity in LSP testing, *Language Testing* 33 (3), 147–152.

Elder, C and Read, J (2015) Post-entry language assessment in Australia, in Read, J (Ed) *Assessing English Proficiency for University Study*, New York: Palgrave Macmillan, 25–46.

European Association for Language Testing and Assessment (EALTA) (2006) *Guidelines for Good Practice in Language Testing and Assessment*, available online: www.ealta.eu.org/guidelines.htm

Fox, J, Wesche, M, Bayliss, D, Cheng, L, Turner, C E and Doe, C (Eds) (2007) *Language Testing Reconsidered,* Ottawa: University of Ottawa Press.

Frost, K and McNamara, T (2018) Language tests, language policy and citizenship, in Tollefson, J W and Perez-Milans, M (Eds) *The Oxford Handbook of Language Assessment Policy and Planning*, Oxford: Oxford University Press, 280–298.

Genesee, F and Upshur, J A (1996) *Classroom-based Evaluation in Second Language Education*, Cambridge: Cambridge University Press.

Harding, L, Alderson, J C and Brunfaut, T (2015) Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles, *Language Testing* 32, 317–336.

Hawkey, R and Milanovic, M (2013) *Cambridge English Exams: The First Hundred Years. A History of English Language Assessment from the University of Cambridge, 1913–2013*, Studies in Language Testing volume 38, Cambridge: UCLES/Cambridge University Press.

Hill, K and McNamara, T (2012) Developing a comprehensive, empirically-based research framework for classroom-based assessment, *Language Testing* 29 (3), 395–420.

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

Hughes, A and Porter, D (Eds) (1983) *Current Developments in Language Testing*, London: Academic Press.

Hughes, A and Porter, D (1984) Editorial, *Language Testing* 1 (1), i–ii.

International Language Testing Association (ILTA) (2000) *ILTA Code of Ethics*, available online: www.iltaonline.com/page/CodeofEthics

International Language Testing Association (ILTA) (2007) *ILTA Guidelines for Practice in English*, available online: www.iltaonline.com/page/ILTAGuidelinesforPra

Isaacs, T, Laurier, M D, Turner, C E and Segalowitz, N (2011) Identifying second language speech tasks and ability levels for successful nurse oral interaction with patients in a minority setting: An instrument development project, *Health Communication* 26, 560–570.

Iwashita, N, Brown, A, McNamara, T and O'Hagan, S (2008) Assessed levels of second language speaking proficiency: How distinct?, *Applied Linguistics* 29 (1), 24–49.

Knoch, U (2011) Investigating the effectiveness of individualized feedback to rating behavior – a longitudinal study, *Language Testing* 28 (2), 179–200.

Knoch, U and Chapelle, C (2018) Validation of rating processes within an argument-based framework, *Language Testing* 35 (4), 477–499.

Knoch, U and Elder, C (2013) A framework for validating post-entry language assessments (PELAs), *Papers in Language Testing and Assessment* 2 (2), 1–19.

Knoch, U and Elder, C (2016) Post-entry English language assessments at university: how diagnostic are they? in Aryadoust, V and Fox, J (Eds) *Current Trends in Language Testing in the Pacific Rim and the Middle East: Policies, Analyses, and Diagnoses*, London: Cambridge Scholars Publishing, 1–20.

Knoch, U and Macqueen, S (2017) Assessment in the L2 classroom, in Loewen, S and Sato, M (Eds) *The Routledge Handbook of Instructed Second Language Acquisition*, London: Routledge.

Knoch, U and Macqueen, S (2019) *Assessing English for Professional Purposes: Language and the Workplace*, Abingdon: Routledge.
Lado, R (1957) *Linguistics Across Cultures: Applied Linguistics for Language Teachers,* Ann Arbor: University of Michigan Press.
Lado, R (1961) *Language Testing: The Construction and Use of Foreign Language Tests – A Teacher's Handbook*, New York: McGraw-Hill.
Lado, R (1964) *Language Teaching: A Scientific Approach,* New York: McGraw-Hill.
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.
McCray, G and Brunfaut, T (2018) Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking, *Language Testing* 35, 51–73.
McKay, P (2005) *Assessing Young Language Learners*, Cambridge: Cambridge University Press.
McNamara, T (1996) *Measuring Second Language Performance*, London/New York: Longman.
McNamara, T (1997) Interaction in second language performance: Whose performance?, *Applied Linguistics* 18 (4), 446–466.
McNamara, T (2001) Ten years of the Language Testing Research Centre, in Elder, C, Brown, A, Grove, E, Hill, K, Iwashita, N, Lumley, T, McNamara, T and O'Loughlin, K (Eds) *Experimenting with Uncertainty: Essays in Honour of Alan Davies*, Studies in Language Testing volume 11, Cambridge: UCLES/Cambridge University Press, 5–10.
McNamara, T and Roever, C (2006) *Language Testing: The Social Dimension*, Oxford: Blackwell Publishing.
McNamara, T and Ryan, K (2011) Fairness vs justice in language testing: The place of English literacy in the Australian Citizenship Test, *Language Assessment Quarterly* 8 (2), 161–178.
McNamara, T, Knoch, U and Fan, J (2019) *Fairness, Justice and Language Assessment*, Oxford: Oxford University Press.
Norris, J M, Brown, J D, Hudson, T and Yoshioka, J (1998) *Designing Second Language Performance Assessments,* Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.
O'Hagan, S (2014) *Variability in Assessor Responses to Undergraduate Essays: An Issue for Assessment Quality in Higher Education*, Bern: Peter Lang.
O'Loughlin, K (2001) *The Equivalence of Direct and Semi-direct Speaking Tests*, Studies in Language Testing volume 13, Cambridge: UCLES/Cambridge University Press.
Pill, J (2016) Drawing on indigenous criteria for more authentic assessment in a specific-purpose language test: Health professionals interacting with patients, *Language Testing* 33 (2), 175–193.
Purpura, J E (2004) *Assessing Grammar*, Cambridge: Cambridge University Press.
Read, J (2000) *Assessing Vocabulary*, Cambridge: Cambridge University Press.
Roever, C and Wigglesworth, G (Eds) (2019) *Social Perspectives on Language Testing*, Frankfurt: Peter Lang.
Shohamy, E (2001) *The Power of Tests: A Critical Perspective on the Uses of Language Tests*, Harlow: Pearson Education.
Shohamy, E and Hornberger, N (Eds) (2008) *Encyclopaedia of Language and Education* (Second edition), Language Testing and Assessment volume 7, New York: Springer Science+Business Media LLC.

Spolsky, B (1995) *Measured Words*, Oxford: Oxford University Press.
Su, L-I W, Weir, C J and Wu, J R W (Eds) (2019) *English Language Proficiency Testing in Asia: A New Paradigm Bridging Global and Local Contexts*, London/New York: Routledge.
Taylor, L (2009) Developing assessment literacy, *Annual Review of Applied Linguistics* 29 (1), 21–36.
Taylor, L (2013) Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections, *Language Testing* 30, 403–412.
Taylor, L B and Chan, S (2015) *IELTS Equivalence Research Project (GM133) – Final Report*, available online: www.gmc-uk.org/-/media/documents/GMC_Final_Report___Main_report__extended____Final___13May2015.pdf_63506590.pdf
Tsagari, D and Banerjee, J (Eds) (2016) *Handbook of Second Language Assessment,* Boston/Berlin: De Gruyter Mouton.
Valette, R M (1968) Evaluating oral and written communication: suggestions for an integrated testing program, *Language Learning* 18, 111–120.
Vernon, P E (1956) *The Measurement of Abilities* (Second edition), London: University of London Press.
Wall, D (2005) *The Impact of High-Stakes Examinations on Classroom Teaching: A Case Study Using Insights from Testing and Innovation Theory*, Studies in Language Testing volume 22, Cambridge: UCLES/Cambridge University Press.
Weigle, S C (2002) *Assessing Writing*, Cambridge: Cambridge University Press.
Weir, C J (1983) *Identifying the language problems of the overseas students in tertiary education in the United Kingdom*, unpublished PhD thesis, University of London.
Weir, C J (1990) *Communicative Language Testing*, New York: Prentice Hall.
Weir, C J (1993) *Understanding and Developing Language Tests*, New York: Prentice Hall.
Weir, C J (2005) *Language Testing and Validation: An Evidence-based Approach*, Basingstoke: Palgrave Macmillan.
Weir, C J and Milanovic, M (2003) (Eds) *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*, Studies in Language Testing volume 15, Cambridge: UCLES/Cambridge University Press.
Weir, C J, Vidaković, I and Galaczi, E D (2013) *Measured Constructs: A History of Cambridge English Language Examinations 1913–2012*, Studies in Language Testing volume 37, Cambridge: UCLES/Cambridge University Press.

Measures of Esteem

This part of the volume contains a series of shorter contributions written by various colleagues and friends of Cyril Weir – some from the early years of his career, some from later years. Contributors were invited to offer something 'closely focused and personal', describing an area of work in which they collaborated with Cyril and what they felt they had learned from that experience. The written contributions contain personalised accounts, memories and reflections from individuals who knew Cyril well and worked with him professionally in different contexts and at different times during his long and distinguished career. These more personal and informal contributions supplement the longer papers in Chapters 1 to 8 to provide professional and personal 'measures of esteem'. They bring to a wider readership appreciative, and sometimes surprising or amusing, insights into what it meant to have Cyril as both valued friend and respected colleague.

1 Working with Professor Cyril Weir: early contacts and long-term collaboration

Roger Hawkey

Former British Council English Language Officer, Asian Institute of Technology Professor, Visiting Professor in English Language Assessment and Evaluation at Centre for Research in English Language Learning and Assessment, Bedfordshire, and Consultant to Cambridge English

The editorial advice to contributors to the Measures of Esteem section of the Cyril Weir memorial volume was that our contributions should be about 'an area we collaborated on with Cyril'. It was also suggested that they should 'be closely focused and personal'. So, here goes.

Cyril Weir and I first became academic colleagues and friends when we were postgraduate students working for our PhDs at the Institute of Education, London University, between 1979 and 1983. Cyril's research focus was the English language demands on students from abroad studying in Britain. His 1983 doctoral dissertation, *Identifying the Language Problems of the Overseas Students in Tertiary Education in the United Kingdom,* proposed an innovative socio-cognitive framework for language test development and validation. This framework was based on his analysis of the English language communication needs of a sample of postgraduate students from abroad. Cyril's dissertation was to prove influential throughout his subsequent career in international English language testing, the field, of course, in which he was to play a leading role and earn a global reputation over more than 30 years.

Like Cyril's, my own concurrent PhD studies (Hawkey 1982) were focusing on the English language and related needs of students from outside Britain studying at British universities. My thesis was that 'multi-dimensional learner profiles can fruitfully inform the design of training programmes for overseas students in the United Kingdom' (Hawkey 1982:iii). My data was collected through a longitudinal study of the factors affecting the success in the UK of a sample of overseas postgraduate students in their doctoral research years here.

So, at the end of the 1970s and the beginning of the 1980s Cyril and I shared key areas of research in the assessment of the current and target English language levels of students for whom English was not their first

language and who were in the UK for their tertiary studies. A main focus, which Cyril and I continued to share over the next three decades, was on the implications of these students' target language levels for their academic performance and other aspects of their lives at their host institutions.

Our academic and personal relationship while we were both doing research in this field was close and proved enduring. So, it would seem appropriate for me to share, in this commemorative volume for Cyril Weir, some of my academic research experiences with him over three decades.

I have decided to focus on a particular research project, on academic reading. This was a project in which Cyril and I played a part along with colleagues with similar applied linguistic and language teaching/testing interests at the Centre for Research in English Language Learning and Assessment (CRELLA) at the University of Bedfordshire. The research concerned is described in a paper entitled: *The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university*. The paper has been published in Studies in Language Testing (SiLT) volume 34 (2012) and in *IELTS Research Reports* 9 (2008). Cyril and I were co-authors with Tony Green, Aylin Ünaldi and Sarojani Devi, all, at the time, colleagues at the University of Bedfordshire. Tony Green is, of course, Cyril's successor as Director of Language Testing Research, Test Development and Assessment Training at CRELLA; Aylin Ünaldi, who gained her second PhD degree at CRELLA, is now Senior Lecturer in TESOL at the University of Huddersfield Department of Education and Community Studies; and Sarojani Devi is a University of Bedfordshire PhD alumna.

Our IELTS academic reading construct, *actual overseas UK-university student reading study,* which is the focus of this contribution, compares the academic reading experience of first-year students at a British university with the reading construct as tested by the IELTS Reading Module. The contextual parameters of the texts read by the target students as part of their study are reviewed and a comparison made with the performance conditions of the reading activities in the IELTS test. The research also investigates the extent to which any problems in reading might increase or decrease according to the IELTS Reading band score obtained before entry to university.

I see this study as typical of Cyril's applied linguistic experience and interests in three ways. Firstly, the research target population is overseas students pursuing their university studies in the UK. Secondly, the key research focus is the assessment of academic reading. Thirdly, the research objective is to assess academic reading with as reasonable face, content, criterion, and construct validity as possible.

The study certainly raises the issue of the relationship between academic reading as a reality for first-year overseas students at a British university and the reading construct of the IELTS Reading test. The major focus of IELTS

is on *careful* reading whereas the data from our survey suggested that for university students' actual academic studies, *expeditious* reading skills and strategies are at least as important, and can also sometimes be more of a problem. This seems to be the case for both L1 and L2 students. However, further research is clearly needed into the comparability of performance on items testing careful and expeditious reading skills and strategies.

Another issue raised by the study is the need to know more about how texts are shaped through the actual test item-writing process. For example, how do the item writers' ideas on the skills being tested through the tasks they set compare with the candidates' protocol reports on them? The study proposes a methodology to identify IELTS texts which have characteristics that may appear *un*typical of actual academic text. Overall, the IELTS texts which were considered as part of this research did appear to fall within the parameters of our small corpus of undergraduate text extracts. But there was some evidence that the demands of even the most 'difficult' of the IELTS texts may actually fall short of those imposed by the most challenging of the academic texts covered in the study.

The final paragraph of the Weir et al study quotes McNamara (1995), who likens performance testing to opening Pandora's box: 'Once it is unlocked, a vast array of questions clamour to be answered, some of which will require detailed intensive study on specific areas' (Weir et al 2012:109). My researcher colleagues and I would certainly agree with McNamara here, recognising, as we do, his relevant research on the use of language tests in immigration and citizenship contexts.

Cyril Weir and I shared careers in the field of applied linguistics, but while his work was UK based with many connections abroad, mine tended to be overseas based, but gravitating home to get qualified and to do my working retirement. Cyril and I met and shared areas of study and research often enough to be long-term good friends. I am very glad about that.

References

Hawkey, R (1982) *An investigation of inter-relationships between personality, cognitive style and language learning strategies: with special reference to a group of adult overseas students using English in their specialist studies in the United Kingdom*, doctoral thesis, Institute of Education, University of London.

McNamara, T (1995) Modelling performance: Opening Pandora's box, *Applied Linguistics* 16 (2), 159–179.

Weir, C J (1983) *Identifying the language problems of the overseas students in tertiary education in the United Kingdom*, unpublished PhD thesis, University of London.

Weir, C J, Hawkey, R, Green, A, Ünaldi, A and Devi, S (2008) *IELTS Research Reports 9: The Relationship Between the Academic Reading Construct as Measured by IELTS and the Reading Experiences of Students in Their First Year of Study at a British University*, Cambridge: British Council/UCLES/IDP.

Weir, C J, Hawkey, R, Green, A, Ünaldi, A and Devi, S (2012) The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university, in Taylor, L and Weir, C J (Eds) *IELTS Collected Papers 2*: *Research in Reading and Listening Assessment*, Studies in Language Testing volume 34, Cambridge: UCLES/Cambridge University Press, 37–119.

2 'The book not written . . .'*

Eddie Williams
CALS, University of Reading, 1977–2003
Department of Linguistics and English Language, Bangor University, 2003–09 (now retired)

'To Eddie,
The book we should have written, but . . .
Best,
Thanks for support,
Cyril'

So reads Cyril's handwritten inscription in the title page of my copy of *Reading in a Second Language* (Urquhart and Weir 1998). At that time Cyril and I had known each other as friends and colleagues since 1986, when he had taken up his post at the Centre for Applied Language Studies (CALS), in the University of Reading. He swiftly made a mark in CALS, initially with his energetic input into the pre-sessional courses. Subsequently, his work in CALS turned to focus on testing and evaluation. He found ready intellectual companions in that field, firstly in the shape of Don Porter and Arthur Hughes, and later with Barry O'Sullivan and his PhD student Tony Green, the last two now respectively Head of Assessment, Research and Development at the British Council, and Director of the Centre for Research in English Language Learning and Assessment (CRELLA) – just two of the many students of Cyril's who have achieved leading positions in the field of testing and evaluation.

In the more than 30 years that I knew him, half of them at CALS, Cyril was an invaluable colleague, both on the academic and personal fronts, always ready to offer advice and constructive criticism. He and I jointly taught a number of modules in the CALS MA programme, and I occasionally took over the Language Testing module when Cyril was engaged in overseas consultancies. Having been securely established by David Wilkins in the mid-1970s, CALS subsequently experienced substantial academic development and physical expansion. In the early years the CALS staff were under self-imposed pressure to balance the books – but we were confident, competent, and happy, and CALS soon established itself as a profitable self-funding unit, under the sound directorship of Ron White, who allowed staff to 'get on with

* With acknowledgement to Robert Frost's *The Road Not Taken* (1916).

it', and to combine the pursuit of their own interests with those of CALS. The result of that enlightened policy was that CALS buzzed with teaching, and the production of books, articles, distance learning modules, research projects, overseas consultancies and conferences – and Cyril was right in the middle of it all. He had a prodigious capacity for hard work, often engaged simultaneously in writing, carrying out research projects and travelling on international consultancies.

As is often the case in the world of academia, Cyril and I did not always see eye to eye on every academic issue, most notably in our approach to the nature of reading in a second language, and consequently to the testing of reading. My experience as a language teacher and a language learner had led me to the view that knowledge of language and linguistic elements plays a crucial role in second language reading comprehension (although of course such knowledge must be complemented by other types of knowledge, and the process is inevitably influenced by contexts of situation and motivation, etc.). Cyril, on the other hand, was inclined to give emphasis to reading as a more global socio-cognitive process, with significant attention to higher-order comprehension, and to 'reading types' such as skimming, scanning, careful reading etc. This is not the place to rehearse these differing viewpoints; suffice to say that our discussions were invariably stimulating and always amicable; however, when the question arose of our co-authoring the above-mentioned book on second language reading, I felt that such disagreements would probably have led to a volume that lacked cohesion and coherence. Alternatively, the endless discussions might have resulted in no book ever appearing, so I withdrew from the plan – there were no hard feelings on either side, and we continued as firm friends.

The first 'Cyril' that I had met in an EFL context was a fictitious character of that name, who, together with his girlfriend Maisie, featured in *Living English Structure* by William Stannard Allen (1947) – a book way ahead of its time, with the introduction saying that 'it does not pretend to tell the student what [they] OUGHT TO SAY in English but tries to show [them] what IS ACTUALLY SAID' (1947:vii, upper case in original). Stannard Allen was not prescient, however, in his depiction of Cyril, with one item in an exercise on tenses, claiming 'Cyril doesn't **often** drink any beer' (1947:112). The student is required to replace 'often' with 'since I first met him', making the necessary verb phrase change. The Cyril that I knew was, without doubt, a true connoisseur of real ale and good pubs. He and I would meet up regularly, along with others, at the Real Ale Festival in London, while his gift for indefatigable research resulted in the discovery of some excellent taverns in London and Berkshire. When frequenting these, we had a rule that no serious decisions of a professional nature were to be made after the first two pints had been consumed, and I commend this rule to everyone. Another favoured ale-quaffing situation was the Six Nations rugby matches, when Cyril would

invite like-minded friends around to his house in Mortimer to watch the televised games – we would drink bottles of Bishops Finger, Abbot Ale and the Reverend James, the religious connotations of these brews perhaps unconsciously signalling the reverence in which Cyril held the game of rugby, and in which he had shown considerable prowess, from his earliest days as a flying winger.

A major feature of the Weir household was his collection of hard-backed books with beautifully illustrated covers, published in the 19th and early 20th centuries. It was an enormous collection, for Cyril rarely did things by halves, while his set of historical novels by the prolific Victorian author G A Henty (1832–1902) must be one of the largest anywhere. These novels featured rousing accounts of English heroism with titles such as *By Sheer Pluck* and *In Freedom's Cause* reflecting Cyril's patriotism, which was pervasive but (perhaps because of his Northern roots in Lancashire) understated. With his customary energy, he had assembled this library largely through rising at the crack of dawn weekend after weekend, and getting to car-boot and bric-a-brac sales before the masses.

This Henty collection, amassed over years, together with his prolific publications, research papers, and conference appearances, demonstrates Cyril's determined and single-minded pursuit of his personal and academic interests. He still found time, however, to help out his many friends and acquaintances when occasion required. Among many acts of kindness, I particularly recall the strong support he provided for our CALS colleague Don McGovern when the latter was struck by cancer, and his solicitous attention to his nephew's academic progress in the tricky transition from sixth form to university, while on the CALS MA programme students such as Peter Davidson, then in Turkey, and Anna Remondi in Brighton were among the many who found Cyril's teaching inspirational. In a different context, when in 2002 I applied for a post in the University of Bangor, Cyril gave me lengthy and meticulous advice on the best way of going forward. His advice must have been pertinent, for I got the job. At that point Cyril and Barry O'Sullivan had already left CALS. As a self-funding unit in a very healthy financial position, CALS had attracted the attention of a revenue-hungry University, obsessed with 'restructuring'; the Centre was eventually merged with the financially challenged Department of Linguistics. Ron White, who had directed the Centre so wisely, had retired. CALS staff were also given to understand that promotion from lecturers (which we all were) to senior lecturers was not on the cards. At this dispiriting news, several staff planned their departure, and within a few years Cyril, Barry O'Sullivan and myself had all left, and become professors in other universities. The boom period of CALS as a virtually independent centre of applied linguistics was waning.

However, in pastures new at the University of Luton, subsequently renamed the University of Bedfordshire, Cyril's boom period was waxing ever

stronger. He became Powdrill Professor in English Language Acquisition, was awarded the OBE, and founded CRELLA, working from a magnificent oak-panelled office, with fine rural views on the Putteridge Bury campus. Although Cyril and I were now working on opposite sides of Britain (Bangor and Bedfordshire), we continued to meet up through conferences, as examiners on PhD *viva voce* examinations, and in occasional lengthy luncheons with other ex-colleagues. There were also, of course, the rugby matches, notably in Cardiff in 2009, when, together with Barry O'Sullivan, Cyril and I saw Ireland beat Wales to win the Six Nations Championship and the Grand Slam. Unable to find accommodation in Cardiff we drove the two hours to Reading immediately after the match, and settled in The Queen's Head to celebrate Ireland and Barry's triumph.

Magnanimity in rugby defeat was another of Cyril's many qualities – but, like all of us, he was not flawless. On occasion he could be irascible, capable of the cutting aside, and inflexible in his views of people and issues, whether academic or political. But, as a friend, I found these blemishes were dwarfed by overwhelming positives of kindliness, good humour and a fierce loyalty to his friends and especially to his family – his wife Shigeko, and children Jamie and Mary. He was a life-enhancing presence whose departure leaves a huge gap in our futures.

References

Stannard Allen, W (1947) *Living English Structure*, London: Longmans, Green and Co.

Urquhart, A H and Weir, C J (1998) *Reading in a Second Language*, London: Longman.

3 Working with Cyril

Jon Roberts

Former lecturer, Centre for Applied Language Studies, University of Reading

Cyril and I were colleagues at the University of Reading from 1987 to 2001 and friends until his premature death in 2018. We collaborated closely for several years: on the Nepal Baseline Study[1] from 1988 to 1991 and the Reading MATEFL *Evaluation of Language Programmes and Projects* module; and as co-authors of *Evaluation in ELT* (1994).

The Overseas Development Administration (ODA) funded the Science Education ELT Project (SEPELT) in Nepal and provided one month's in-service training (INSET) for 1,080 secondary-level teachers of English with the aim of raising student performance in the National School Leaving Certificate. It was delivered by locally trained Nepali staff supported by an expatriate training officer and ran from 1987 to 1989. Subject to the Thatcher government's pursuit of value for money ('the optimal use of resources to achieve intended outcomes', according to the National Audit Office[2]), the Education Division of ODA decided on a baseline study design for the evaluation of SEPELT. Its purpose was to measure the impact of SEPELT on the English language performance of students and to assess the efficacy of baseline design for future project evaluations (Department for International Development 1991). While a standard and widely reported design for accountability-oriented evaluation in the US, baseline project evaluation was a relatively new approach in the UK and a first in ELT.

Cyril's role was to plan the baseline design (choosing a small-scale, non-equivalent control group pre-/post-testing study); to develop a test battery on teachers' language levels and the performance of learners taught by trained and untrained teachers; to recruit and train local technical staff as data collectors; and to analyse and interpret the data. He invited me to work on observational tools to assess differences, if any, in the classroom behaviour of trained and untrained teachers. This required extensive planning in the UK, visits to schools and training sessions in central and southern Nepal to develop and monitor data collection, and data analysis and presentation.

As the junior partner, I learned a lot from Cyril. He set a standard for energy and drive, attention to detail within a clearly conceived bigger picture, meeting deadlines, and expertise. This was combined with an ability to work well with people at every stage and level in the project: he enjoyed hard-nosed negotiations with British and Nepali administrators perhaps just as much as

hanging out after hours with our Nepali informants and field workers (some of whom, tellingly, became friends).

Collaboration can be beneficial because it uncovers differing points of view and ours helped develop my own thinking. Cyril was interested in group-level, quantifiable hard data, while I was more interested in the perspectives of individual teachers. I will never forget the teachers attending an INSET session in Pokhara who had walked from the other side of the Himalayas and were in desperate need of their lunch of two chapatis and an egg. They taught huge classes with few books and with relatively little impact on student learning.[3] One had to admire their wish for efficacy and ask how to understand what their priorities were, how far they coincided with the bureaucracy, and how best to use INSET budgets. The literature on expensive curriculum innovations was full of reports of a gap between individual teachers and a new curriculum, and combined with experiences in Nepal and elsewhere, I was led to greater interest in INSET for individual teacher development and theory adequate to understand it (see, for example, Roberts 1998).

Literally and metaphorically Cyril and I ended up in very different places in our work, but the Nepal experience showed us how technically demanding and socio-politically fascinating programme evaluation was; and how essential systematic evaluation was for accountability and continuous improvement. At that time there was a gap in the ELT literature and so we decided to produce a book on evaluation, illustrated by case studies and supported by a menu of data collection methods (Weir and Roberts 1994). We planned the structure, co-wrote the introduction, divided up the other chapters and edited each other's work face to face and in detail.

The merits of the book are for others to say and it has been out of print for a long time, but I can comment on the professional benefit of a year and a half of co-authorship. It was an apprenticeship in making a book-length text and enjoying it: the pleasure of building sentences into paragraphs, paragraphs into chapters, and chapters to make a whole. I learned the importance of contents pages; of trying to take the reader with you by careful signposting; the necessary and often trying recursiveness of writing as you cross-refer separated sections for repetition, redundancy and self-contradiction; the painful pleasure of cutting words, phrases and even whole paragraphs – 'killing your children'; and of course the pleasure of giving a box-fresh hardback to Mum and Dad. Less obviously I learned the underrated pleasure of picking out the key content of sub-sections to build up an index (which often led to further editing) and using it to survey the content of the whole book. Cyril was particularly, one might say forensically, strong on coherence and relevance. He thought in paragraphs and was ruthless in rooting out gaps, woolliness, redundancies and inconsistencies. And most of all he made it fun. He hugely enjoyed writing, editing and collaborating: getting it right with gusto.

When Essex meets Birkenhead, co-writing becomes a raucous, occasionally juvenile business and we evolved our own largely unprintable editorial shorthand. It was fun. Re-reading even this short piece as I am now, I can imagine Cyril trawling for mistakes and bad writing (the 'shit detector'); hunting down self-contradictions and unsupported claims ('arse covering'); and, awarding the ultimate praise for a coherent story told in economic prose: 'shit off a shovel' and time for a drink!

References

Department for International Development (1991) *The Role and Design Of Baseline Studies In The Evaluation Of English Language Training In The Case Of Nepal (ev485)*, London: Department for International Development.
Roberts, J (1998) *Language Teacher Education*, London: Arnold.
Weir, C J and Roberts, J (1994) *Evaluation in ELT*, Oxford: Blackwell.

Endnotes

1 'A **baseline study** is an analysis of the current situation to identify the starting points for a program or project. It looks at what information must be considered and analyzed to establish a **baseline** or starting point, the benchmark against which future progress can be assessed or comparisons made.' (Eurosat, ec.europa.eu/eurostat/statistics-explained/index.php/Glossary: Baseline_study; emphases in original)
2 www.nao.org.uk/successful-commissioning/general-principles/value-for-money/assessing-value-for-money/
3 Schools in urban centres such as Bharatpur and Pokhara were much better resourced.

4 Reflections from Egypt: the role of Cyril Weir in national assessment reform initiatives

Hanan Khalifa
Cambridge Assessment English

Writing reflections on an influential figure like Professor Cyril Weir could seem like a huge undertaking, but writing about Cyril, my mentor, supervisor and friend, evoked many happy memories and made me reflect on how to continue learning from his teaching, his thinking and his impact. In the next few paragraphs I would like to share the impact Cyril had on English language assessment reform initiatives in my home country of Egypt. Specifically, I will reflect on the technical assistance he provided to initiatives in higher education and in the K-12 sectors between 1991 and 2003.

Higher education

Between the 1970s and the 1980s, the UK Overseas Development Administration (the ODA, now known as the Department for International Development) helped in promoting English for Specific Purposes (ESP) in the Middle East, especially in countries where universities started using English as a medium of instruction. One of those countries was Egypt, and more specifically, Alexandria University, where the first ESP Center in the Middle East was established.

The strategic objectives of the ESP Center were twofold. Firstly, the Center activities should enable undergraduate students across Alexandria University to access their subject specialisation through ESP teaching, curricula and materials; and secondly, it should assess students' English language proficiency for admission onto postgraduate courses and degrees. Professor Cyril Weir had a long-lasting impact in relation to the second of these objectives which I will detail below in chronological order.

- From 1991 to 1992, Cyril was contracted by the ODA to build the capacity of a core team at the ESP Center in order to develop and validate the Alexandria University EAP test battery. The team learned from Cyril key principles of assessment, gained an understanding of the intricacies of national and international tests (Test of English for Educational Purposes (TEEP), International English Language Testing

System (IELTS), Test of English as a Foreign Language (TOEFL)), and became aware of the importance of considering practical issues when designing an assessment tool.

- Between 1992 and 1994, Cyril influenced the establishment of the first ever testing and evaluation unit in an Egyptian University, and perhaps in the Middle East. To him, it was the cornerstone for continually improving assessment, and for the Center it became a milestone in reforming English language assessment provision in higher education. He also led the development and validation of the English for Academic Purposes (EAP) test battery which continues to be in use at the time of writing these reflections. During this time, Cyril completed his 1993 book *Understanding and Developing Language Tests* (published by Prentice Hall). The work he did with the Alexandria team helped refine his thinking on test operations and conditions, and the team received warm words in his acknowledgments in the book.
- From 1994 to 1998, Cyril acted as a supervisor, mentor and coach to four team members of the testing and evaluation unit who went on to gain MA and PhD degrees in Language Testing. The four of us (plus Dr Kamal El Fouly, a student of Professor Lyle Bachman at University of California, Los Angeles) were the very first batch of language testers in Egypt to receive internationally recognised degrees in language testing. It was through his mentorship, guidance and example that we went on to influence thinking on assessment in the Egyptian higher education sector and to raise awareness of the importance of assessment literacy for all stakeholders.

K-12 sector

In 1997, the United States Agency for International Development (USAID) launched its programme to support Egypt in reforming its English language provision in the K-12 state school sector. A key component of this programme was reforming the state examination system via four strategic priorities. These were:

1. Fostering organisational collaboration and dissemination of information on testing.
2. Developing a quality instrument to assess the language proficiency of Egyptian English language professionals.
3. Developing quality student achievement tests based on the curriculum and textbooks in use.
4. Provision of in-service training courses in test design and classroom assessment.

Developing a national capacity to design, administer and analyse English language measurement and evaluation instruments was integral to the realisation of the above priorities and essential to the sustainability of the test reform undertaking.

From 1998 to 2003, I was responsible for the planning and execution of the assessment reform programme. As such, I sought Cyril's support in realising strategic priorities 3 and 4, not only because he was a known figure in the field who ensured greater credibility regarding programme outputs, but also because he had a unique ability to combine academic knowledge with financial awareness – a capability needed to ensure return on investment and sustainable development. The programme had several outputs and I would like to highlight a few below where Cyril had the most impact through his foresight and guidance:

- *Creation of a trained cadre at a national level.* The cadre was carefully selected with sustainability in mind and procedures were set in place to guard against attrition.
- *Official recognition of the cadre as language testing specialists.* This was a critical step for sustainability purposes.
- *Introduction of language testing topics to the annual training plan for in-service teachers.* Training was conducted by the trained cadre at a local level, which reduced costs and ensured economic efficiency during the budget planning phase.
- *Production of a student achievement test development manual.* The manual provided a step-by-step guide to textbook analyses for assessment purposes, test construction, item writing development and moderation, test administration and test evaluation. Copies of the manual were produced locally and distributed to schools and districts.
- *Improved key stage test specifications and introduction of listening and speaking in school-based assessment.*

When working with Cyril on the programme, he always brought in the international perspective while being sensitive to and showing understanding of the local context. Through his logical arguments, he was instrumental in helping me to persuade key stakeholders that the policy of cherry picking may not lead to intended consequences and that a 'one size fits all' mentality may have adverse impact on reform initiatives. The trained cadre which he helped establish always referred to him as the 'firm, friendly and fair Cyril'.

The above are merely two examples of how Cyril has impacted on individual growth, institutional capacity building and local national initiatives. There are many more examples worldwide attesting to how influential he was in the field of language testing and in assessment reform.

5 Reflections from Taiwan: the contributions of Cyril Weir to the GEPT and the glocalisation of English language proficiency testing in Asia

Jessica R W Wu
Language Training and Testing Center, Taiwan

To acknowledge Cyril Weir's impact and influence on the General English Proficiency Test (GEPT) and the Language Training and Testing Center (LTTC) is easy. However, to pin down his contributions in prose is a much more difficult task, for they are enduring and pervasive. Cyril joined the GEPT Research Committee in 2001 and continued to serve as an external consultant thereafter. Thanks to his invaluable guidance and continuous support, the GEPT has thrived in the two decades since then. Indeed, in 2020, the GEPT is going to celebrate its 20th anniversary, a truly extraordinary milestone for a locally produced English proficiency test and an achievement which would not be possible without Cyril's visionary mentoring.

Through his extensive experience of working with testing developers around the world, Cyril understood the necessity of test localisation and therefore supported the creation of the GEPT. This led to the LTTC realising its plan to provide a localised English proficiency test in Taiwan as an alternative to international standardised tests such as IELTS and TOEFL. Through the combined forces of the Testing Research Committee, which consisted of international testing experts, including Cyril, and the Test Advisory Committee, which consisted of scholars and academics in Taiwan, the GEPT was tailored specifically to the local learning context and aligned with local curricula.

On the subject of test localisation, Cyril's attitude was very clear: he repeatedly stressed that test quality cannot be compromised under any circumstances. Two memorable illustrations of how his convictions helped shape the GEPT spring to mind.

The first was when Cyril paid his very first visit to Taiwan as part of his consultancy during the design and development of the GEPT advanced and superior levels. I remember at that time my colleagues and I sitting with him in a conference room and feeling rather overwhelmed by his presence, for we

were not sure whether we would be able to deliver such a challenging and innovative test in Taiwan. It felt very much a mission impossible! He nonetheless managed to convince us that in order to be the leader in language testing in Taiwan and shift the public's perception of language testing, the mission had to be completed no matter how impossible it might initially appear. The advanced-level test was intended for English for Academic Purposes (EAP), so it was vital that the test construct and tasks reflected real-life tasks in order to assess and elicit higher-level cognition processes from test takers. We soon realised that the multiple-choice response format, which was considered of great practicality for a large-scale standardised test, could not achieve this. The short-response format had to be incorporated. So this is what we did. It felt like quite a bold move, for there were many administrative difficulties to be overcome, not to mention the fear of scaring test takers away by presenting them with an unfamiliar response format. Time and validation studies, as I will elaborate later, have proved Cyril right, and I am still grateful to him to this day for his insistence on this point.

To highlight Cyril's foresight of more than a decade ago, I must mention the rationale he provided us with to support the GEPT advanced level. The reading test consists of two parts which assess competence in two different reading strategies: expeditious reading and careful reading, in other words, reading for gist and specific information versus reading for detail and authorial stance. Again, Cyril's work had great bearing on the LTTC's approach. As his research has long argued, reading can be viewed as comprised of multi-divisible skills, rather than as a unitary skill (Weir and Porter 1994). In real life, people employ different reading strategies for different purposes, and the test construct should reflect this. Informed by this understanding, the GEPT advanced reading test is one of few tests that separate these distinct reading strategies into two components, thus being able to assess intertextual comprehension and interpretation. Another attempt to make the GEPT advanced more closely resemble real-life tasks can be found in the writing test. The way the LTTC incorporated this insight was to split the writing test into two parts. In the first part, verbal input is provided. Test takers have to assimilate two 400-word texts that represent two opposite views on the same issue, summarise these two texts, and then express their personal opinions on the matter. In the second part, non-verbal input is provided. Test takers have to read two charts or graphs, summarise the main findings from them, and then provide solutions or suggestions. The integrated reading-into-writing task type and the testing of two reading strategies are state-of-the-art features inherent in the GEPT advanced test. Both are the fruits of Cyril's inspiration.

The second area where Cyril's influence helped mould the LTTC's endeavours is in the field of validation. Over the years, as the stakes of the GEPT have continued to rise, the responsibility of ensuring quality has fallen solely on us and has become the impetus for us to do better. All GEPT validation

studies have been based on Cyril's socio-cognitive framework (Weir 2005), and some in fruitful collaboration with Centre for Research in English Language Learning and Assessment (CRELLA). His groundbreaking framework systematically identifies the evidence required to establish a comprehensive and coherent validity argument, while simultaneously investigating the interaction between different types of validity evidence. His profound insights have provided a reference and scaffold to guide the data collection process and research agendas, all aiming to establish the validity of the GEPT in its various aspects. Thanks to these studies (see selected studies listed in Table 1; a complete list of GEPT research is available at www.lttc.ntu.edu.tw/thesis.htm), conducted jointly by the LTTC's in-house research team and external researchers, including, of course, Cyril Weir himself, the GEPT accords with international standards. Of the research listed in the table, my colleague Rachel Wu's 2011 study of establishing test validity through examining alignment with the CEFR deserves special mention. It was not conducted by Cyril himself, but it was supervised closely by him. Rachel had the privilege of being mentored by Cyril while pursuing her PhD at the University of Bedfordshire. Her doctoral thesis has been considered as a good example in its area, and it was published jointly by Cambridge Assessment English and Cambridge University Press in 2014.

All of the studies have provided significant evidence in terms of validating the GEPT. I am certain that without Cyril's initiative and supervision, our efforts would not have been as well developed as they currently are. As a result, the GEPT has gained wide recognition by leading academic institutions both at home and abroad and has drawn interest from researchers around the world (see www.gept.org.tw/ORG/gept_02_03_list.asp). This is a significant accomplishment for a locally produced English proficiency test in Asia. Every time I reflect on what the GEPT has achieved, I realise how indebted we are to Cyril for guiding us through the journey.

In addition to his work in Taiwan, Cyril also collaborated closely with other examination boards in Asia, including College English Test (CET) in China and EIKEN in Japan. He took all of us under his wing, offering advice and constructive criticism. He helped us believe that although we are not English native speakers, we can nevertheless produce tests that stand up to any produced by native speakers. As one of Asia's most prestigious testing organisations, we at the LTTC feel obligated to share our expertise with fellow testing organisations so that we can carry on his legacy. Only through sharing and collaboration can all in the language testing family continue to grow and excel. This is especially true for Asian countries, since we share similar educational systems and learning environments.

Witnessing how the locally produced English proficiency tests in Asia have shifted the landscape of English language testing and learning, Cyril and I felt the time was ripe for the publication of a book which focused solely on

these local tests and provided an insightful overview of them from the test development perspective. *English Language Proficiency Testing in Asia: A New Paradigm Bridging Global and Local Contexts* (Su, Weir and Wu (Eds) 2019) is that very work, and will be a fine testament to him. The extent of the faith he had in us can be seen in his refutation of the prevailing assumption that international tests, developed by English native speakers and with a longer history, are superior to local ones (see Weir 2013). Acknowledging that the trend in Asia towards localising English proficiency tests was likely to continue, he believed the six tests introduced in this new volume would serve as a model for the assessment not only of English but also of other foreign languages for test developers all over the world.

Cyril contributed the concluding chapter to this book as well as editing other contributions shortly before he passed away. It is a great sadness to me that he was not able to see *English Language Proficiency Testing in Asia: A New Paradigm Bridging Global and Local Contexts* published; however, the work serves as a token of the gratitude and respect felt towards Cyril by all of its contributors. I believe that all of us who worked on Cyril's final book regard it as a great honour to continue his quest to make a positive impact on English learning and teaching, and on society as a whole, through testing.

I would like to conclude this article on a personal note. It was Cyril who brought me into the academic world of language testing and helped me to thrive in my career even after I completed my doctoral study under his supervision in 2005. It is no exaggeration to say it was Cyril who took me from 'crayons to perfume' as the song *To Sir with Love* has it. Indeed, I remember fondly the times when he enjoyed pitchers of margarita with me and my colleagues on the GEPT team during his visits to Taipei. The memory of him sitting among us, sharing in our delight of the progress we were all making on the GEPT, lives on, as does Cyril's legacy.

Table 1: Selected GEPT validation studies

Type of validity evidence	Study
Context and cognitive	• Chan, S H C, Wu, R Y F and Weir, C J (2014) *Examining the context and cognitive validity of the GEPT Advanced Writing Task 1: A comparison with real-life academic writing tasks*, LTTC–GEPT Research Report No. RG-03, Taipei: The Language Training and Testing Center. • Bax, S and Chan, S H C (2016) *Researching the cognitive validity of GEPT high-intermediate and advanced reading: An eye tracking and stimulated recall study*, LTTC–GEPT Research Report No. RG-07, Taipei: The Language Training and Testing Center.

Table 1: Selected GEPT validation studies (*continued*)

Type of validity evidence	Study
Scoring	• Weir, C J and Wu, J (2006) Establishing test form and individual task comparability: A case study of a semi-direct speaking test, *Language Testing* 23, 167–197. • Wu, R Y F (2016) Creating a common score scale for the GEPT to support interpretation of learning progress, in Leung, Y N (Ed) *Epoch Making in English Language Teaching and Learning*, Taipei: Crane Publishing, 223–236.
Consequential	• Wu, J and Lee, M (2017) The relationships between test performance and students' perceptions of learning motivation, test value, and test anxiety in the context of the English benchmark requirement for graduation in Taiwan's universities, *Language Testing in Asia* 7. DOI: 10.1186/s40468-017-0041-4
Criterion: CEFR linking studies	• Brunfaut, T and Harding, L (2014) *Linking the GEPT listening test to the Common European Framework of Reference*, LTTC–GEPT Research Report No. RG-05, Taipei: The Language Training and Testing Center. • Knoch, U (2016) *Linking the GEPT writing sub-test to the Common European Framework of Reference (CEFR)*, LTTC–GEPT Research Report No. RG-08, Taipei: The Language Training and Testing Center.
Criterion: Comparison studies with other CEFR-aligned tests	• Wu, R Y F (2011) *Establishing the validity of the General English Proficiency Test reading component through a critical evaluation on alignment with the Common European Framework of Reference*, unpublished doctoral dissertation, University of Bedfordshire. • Weir, C, J Chan, S H C and Nakatsuhara, F (2013) *Examining the criterion-related validity of the GEPT advanced reading and writing tests: Comparing GEPT with IELTS and real-life academic performance*, LTTC–GEPT Research Report No. RG-01, Taipei: The Language Training and Testing Center. • Wu, R Y F (2014) *Validating Second Language Reading Examinations: Establishing the Validity of the GEPT Through Alignment with the Common European Framework of Reference*, Studies in Language Testing volume 41, Cambridge: UCLES/ Cambridge University Press. • Kunnan, A and Carr, N (2015) *Comparability study between the General English Proficiency Test–Advanced and the Internet-Based Test of English as a Foreign Language*, LTTC–GEPT Research Report No. RG-06, Taipei: The Language Training and Testing Center.

References

Su, L I-W, Weir, C J and Wu, J R W (Eds) (2019) *English Language Proficiency Testing in Asia: A New Paradigm Bridging Global and Local Contexts*, London/New York: Routledge.

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

Weir, C J (2013) *The Way of Language* 1, 6–7.

Weir, C J and Porter, D (1994) The multi-divisible or unitary nature of reading: The language tester between Scylla and Charybdis, *Reading in a Foreign Language* 10 (2), 1–19.

6 Travels with Cyril

Lynda Taylor
Centre for Research in English Language Learning and Assessment, Bedfordshire

Cyril was well known and highly regarded for many things in his professional life; among other things, he was a great travelling companion on professional trips overseas. This short piece shares a few memories of my own time 'on the road with Cyril' as we travelled together to teach and lecture on courses in language testing and assessment in various parts of the world over the past decade or so.

Just over 10 years ago, Cyril and I began co-leading one of the regular annual training courses organised by the Association of Language Testers in Europe (ALTE). In previous years he had shared ALTE course leading with Barry O'Sullivan and with Hanan Khalifa, in Perugia, Munich and Valencia, but in 2008 the course took place in the beautiful city of Prague in the Czech Republic, hosted by the ALTE member there, with Cyril and myself as course leaders.

Cyril sometimes mused that in another life he would have been an antiquarian and second-hand bookseller and anyone who saw all the bookshelves in his garage would probably agree! But Cyril might have had an equally successful career as an upmarket and niche tour operator, for he had an uncanny knack of identifying bijou hotels in all the European cities we travelled to together. The hotels he picked were almost always above the accommodation budget limits set by those who held the ALTE course purse-strings, but Cyril could often negotiate a good package direct with the hotel and so get exactly what he wanted. And I was always happy to tag along . . .

In September 2008 we stayed in central Prague. We would teach sessions each day from about 9.30 a.m. to 4 p.m. and then enjoy the local sights and a meal together in the evening, sometimes in the company of the other course participants. 2008 was also the year when Cyril ruptured his Achilles tendon and, though he was no longer wearing an air boot by September, walking alongside him over the cobblestones in the streets of Prague caused me considerable anxiety all week in case he should fall and need to be hospitalised. I wasn't sure how I would get him safely back to the UK and then explain it all to his wife, Shigeko!

Exactly a year later we spent a week co-teaching the ALTE course in Venice, a city which really took Cyril's fancy and to which he later returned with the whole family. This time, walking round the alleyways and bridges of Venice was an absolute joy, as were our trips on the *vaporetto* water taxis and

the glass of Prosecco with our sandwich at lunchtime. I tried hard to persuade Cyril to take a gondola with me on the Grand Canal, but he said he'd leave that pleasure to my husband!

September 2010 saw us in Bilbao in Spain's Basque Country. This time Cyril chose a city centre hotel right opposite the Guggenheim Museum. Each morning at breakfast we sat on the restaurant balcony overlooking Frank Gehry's architectural wonder clad in titanium and glass, with its flowing curves and reflective surfaces. Cyril and I were less impressed by the exhibition of works inside the museum by British sculptor Anish Kapoor, including a cannon that fired large blood-red pellets of wax onto snow-white walls in one corner of the gallery. The museum brochure described this exhibit as 'a work reflecting the artist's interest in the self-made object; as the wax builds up on the walls and floor of the gallery the work slowly oozes out its form'. Sadly, I can't really repeat here Cyril's verbal evaluation of this particular piece of artwork but those who knew Cyril well can imagine the colourful description he offered!

A year later, in 2011, we moved from southern Europe to its northern shores – to Copenhagen in Denmark. This time our hotel accommodation was in a heritage property – the Admiral Hotel – converted from ancient harbour wharf buildings dating back to the 1780s. We were fortunate to be staying only 5 minutes' walk from Nyhavn where at the end of the day you could sit outside with a leisurely drink in the waterfront cafés and watch the boats coming and going. Cyril was shocked, however, at the local price of a gin and tonic – our routine beverage after a day's teaching together. Always one to find a solution to a problem, Cyril tracked down a local supermarket where we purchased our own bottle of gin, our own supply of tonic, and even a lemon! I'm embarrassed to say that there is a room in the Admiral Hotel in Copenhagen where the marble surface in the bathroom has been irreparably bleached by Cyril's repeated cutting of slices of lemon for our gin and tonics over the five days of our stay!

The ALTE course in 2012 didn't involve any overseas travel as it was based in Cambridge. I stayed in my own home that year but picked Cyril up every morning from the Varsity Hotel in Cambridge where he had once again found himself an attractive accommodation package. After breakfast he would fold himself into the front seat of my little silver Mini Cooper to travel to Hughes Hall for the day's sessions; and at the end of the day we would debrief and relax together high up on the Varsity's roof garden with its spectacular 360-degree panoramic view of Cambridge.

2013 saw us in Sofia, Bulgaria, for a week and this time I brought along my recently retired husband to make it a threesome. There were distinct advantages to having Nigel with us for the week as I had never been able to do justice to the bottle of red wine that Cyril would routinely order with our evening meals on the ALTE course. As a long-standing Francophile, Nigel

was well placed to share a bottle (or bottles) of fine red wine with Cyril and to talk about education, rugby and football, especially Liverpool's ups and downs.

The following year, 2014, the three of us were together again, this time in Paris – more red wine and more fine dining! In the suburb of Sèvres we found a bijou restaurant just around the corner from the hotel which could only accommodate about a dozen diners and which offered a fresh menu every day. The cuisine was so good that we decided to eat there three nights running and were never disappointed.

Paris 2014 was our last ALTE course together and my Septembers were never to be the same again.

As I reflect back, I realise what a good friend and travelling companion Cyril could be over a week in a European city – whether it was chatting about work or family, history or politics, art or culture, or so many other things. I will always cherish the professional and personal friendship we enjoyed together, which had time to grow and blossom on our teaching trips across Europe.

And within that friendship, I want to acknowledge Cyril's faithful support and encouragement of me as a female academic – one who came fairly late to the academy. Cyril gave me research and editorial opportunities that I could never have dreamed of in my early career as an EFL teacher. I suspect there are several of us female academics who will testify to the confidence Cyril placed in us as students and researchers and the encouragement and support he gave us in our professional lives.

I also came fairly late to the role of priest in the Anglican Church. When I was growing up there was no question of women becoming priests and certainly no role models to look to; that didn't change until 1994 and even female bishops have only been with us in the Church of England since January 2015. I shall always be grateful for Cyril's support for and encouragement of me in my other profession – as a female priest. He would occasionally give me a book on theology or a collection of 19th century sermons – usually discovered on his regular visits to local second-hand bookshops or car boot sales.

In 2013, around Christmas time, Cyril gave me a copy of Eerdmans *Dictionary of the Bible* – an excellent resource that I often refer to when preparing for a church service. In the front of the book Cyril wrote the following words – in that utterly unique spidery handwriting of his:

> *To Lynda – How to be smarter than the average bishop – sapientia est potentia (wisdom is power) – even if it comes from a car boot sale bargain. Best, Cyril.*

Cyril was indeed for me 'the best'.

Appendix 1
Obituary: Professor Cyril J Weir (1950–2018)

Anthony Green

Centre for Research in English Language Learning and Assessment, Bedfordshire

It is with deep sadness that we announce the death on September 28 of Cyril J Weir, the Powdrill Research Professor in English Language Acquisition at the Centre for Research in English Language Learning and Assessment (CRELLA) at the University of Bedfordshire.

Born on Merseyside in 1950, Cyril pursued his early enthusiasm for radical politics through an undergraduate degree at the University of Reading, followed by a Certificate in Education in Liberal Studies from the University of Birmingham in 1971–72. He lectured in European Studies at Middlesex Polytechnic while studying for his Master's in Political Philosophy at Reading. Cyril first became involved in the world of English language education when he set out to fund further studies in History by working as a lecturer in EFL in Iran. Finding a new intellectual direction, he signed up for the University of Edinburgh course in Applied Linguistics and encountered the world of Language Testing in the inspiring shape of Alan Davies.

After leaving Iran in the late 1970s as the stirrings of the 1979 revolution took hold, Cyril found a position as a Research Officer at the Associated Examining Board (AEB, a UK-based examinations provider that is now part of AQA). He was tasked with developing a new test to screen international students entering UK universities. Adapting John Munby's (1978) approach to needs analysis, Cyril carried out a comprehensive programme of observation and interviews to reveal how university students used language and their major sources of difficulty. This was the basis of his PhD (1983) from the Institute of Education (University of London) under the supervision of Chris Brumfit, and led to the AEB Test in English for Academic Purposes (TEAP). Both the original development process and innovative features of the test, such as the use of tasks that integrated reading, listening, and writing skills, remain influential 35 years later. A revised version of the test itself remains in use at the University of Reading as the Test in English for Educational Purposes (TEEP).

Although he returned to university teaching in 1983, Cyril was always

keen that his ideas should serve practical ends. He found like-minded colleagues at the Centre for Applied Language Studies (CALS) at the University of Reading where he teamed up with Arthur Hughes, Don Porter and others to build the Testing and Evaluation Unit (TEU), which he went on to direct from 1996. His books *Communicative Language Testing* (1990) and *Understanding and Developing Language Tests* (1993) together represented the most coherent case made for the needs-based approach to language test development that had emerged in the UK over the previous decade. *Evaluation in ELT* (with Jon Roberts) (1994) and *Reading in a Second Language: Process, Product and Practice* (with Sandy Urquhart) (1998) cemented his reputation.

At Reading he directed and led numerous test and evaluation projects, including acting as the Senior UK Consultant on the College English Test and Test for English Majors projects in China, the Universities' EAP Proficiency Test Project in Egypt, and on UK Overseas Development Administration (ODA) evaluation studies in Nepal, Guinea and Ecuador. Clients and students alike appreciated his ability to combine his sharp insights with a clear sense of the steps required to put ideas into practice. In addition to teaching at Reading, he built up a collaborative relationship with the Association of Language Testers in Europe (ALTE), working as a senior consultant and trainer. The annual Summer Course on Language Assessment that he developed at Reading in 1996, working with Barry O'Sullivan and Rita Green, was adopted as the ALTE Summer Course. Cyril served as the lead presenter from its inception in 2005 until 2016.

Taking up a Professorship in ELT at the University of Surrey, Roehampton in 2000, Cyril developed his influential socio-cognitive framework for test development and validation, which became the centrepiece for *Language Testing and Validation: An Evidence-based Approach*, published in 2005 (Weir 2005). The framework combines social, cognitive, and evaluative (scoring) dimensions of language use and links these to the context and consequences of test use. The framework shaped the work of testing organisations both in the UK (e.g. Cambridge Assessment English and the British Council) and internationally (e.g. the Language Training and Testing Center (LTTC) in Taiwan and Shanghai Jiao Tong University, where Cyril was appointed as Visiting Professor). Cyril became joint Series Editor for the Cambridge Assessment English/Cambridge University Press Studies in Language Testing (SiLT) series (first with Michael Milanovic and then Nick Saville) and further elaborated the socio-cognitive framework in the SiLT volumes *Examining Writing* (with Stuart Shaw) (2007) and *Examining Reading* (with Hanan Khalifa) (2009). In the latter part of his career, Cyril found a place for his earlier passion for history, publishing two volumes on the evolution of English language testing: *Measured Constructs: A History of Cambridge English Language Examinations 1913–2012* (with Ivana Vidaković and

Evelina Galaczi) (2013) and, for the British Council, *Assessing English on the Global Stage: The British Council and English Language Testing 1941–2016* (with Barry O'Sullivan) (2017).

As Powdrill Chair in Second Language Acquisition at the University of Bedfordshire from 2005, Cyril established the Centre for Research in English Language Learning and Assessment (CRELLA). As a specialist research centre with a clear focus on language assessment, CRELLA grew and thrived under his 10-year directorship, winning national and international recognition as a world-leading centre of excellence. His many contributions were recognised with a Distinguished Achievement Award from Cambridge/ILTA (International Language Testing Association) in 2014, election as a Fellow of the Academy of Social Sciences in 2013, and in 2015, for his services to English language assessment, an OBE (Officer of the Most Excellent Order of the British Empire, a UK national honour for individuals playing an important public role).

Cyril mentored and supervised many language testers, who have gone on to become leaders within our field across the globe. The many of us who had the privilege to work or study with Cyril over his 40-year career, will miss his humour and his insightful, often critical, but always practical advice, delivered with warmth and generosity of spirit.

Cyril leaves a wife, Shigeko, and two children.

References

Khalifa, H and Weir, C J (2009) *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing volume 29, Cambridge: UCLES/Cambridge University Press.

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

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

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

Weir, C J (1983) *Identifying the language problems of the overseas students in tertiary education in the United Kingdom*, unpublished PhD thesis, University of London.

Weir, C J (1990) *Communicative Language Testing*, New York: Prentice Hall.

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

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

Weir, C J and O'Sullivan, B (2017) *Assessing English on the Global Stage: The British Council and English Language Testing 1941–2016*, Sheffield: Equinox.

Weir, C J and Roberts, J (1994) *Evaluation in ELT*, Oxford: Blackwell.

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

This obituary first appeared in *Language Assessment Quarterly* 15(4), 25 October 2018.

Appendix 2
Curriculum vitae of Professor Cyril J Weir, AcSS MA MSc PhD

Qualifications

1968–71 University of Reading, BA (Iii) Politics
1971–72 University of Birmingham, Cert Ed Liberal Studies
1972–74 University of Reading, MA Political Philosophy
1977–78 University of Edinburgh, MSc Applied Linguistics
1979–83 Institute of Education, University of London, PhD Language Testing

Appointments

1972–74
Lecturer in European Studies, Middlesex Polytechnic
1975–77
Senior Lecturer in EFL, MIS Technical College Iran
1979–83
Research Officer, Associated Examining Board, Aldershot
1983–85
Lecturer in ELT, University of Exeter
1985–86
Teaching Fellow in ELT, University of Lancaster
1986–2000
Lecturer in EFL, Director of Testing and Evaluation Unit, Centre for Applied Language Studies (CALS), University of Reading
2000–2005
Professor in ELT, University of Surrey, Roehampton
2006–18
Visiting Professor, Shanghai Jiao Tong University
2005–18
Powdrill Professor in English Language Acquisition, Director of Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire

Positions of responsibility held

UK
1983–86
Chief Examiner in English as a Foreign Language, Institute of Linguists
1983–91
Chief Examiner in Test in English for Educational Purposes, Associated Examining Board, Aldershot
1988–91
Language Advisor to General Medical Council PLAB test for overseas medical professionals

Overseas

1988–93
Senior Consultant, Evaluation Department, Overseas Development Administration, including baseline evaluation studies of language projects in Nepal, Guinea and Ecuador
1990–95
Co-ordinator, Universities' EAP Proficiency Test Project, Egypt
1991–95
Senior UK Consultant, College English Test (CET) Validation Study, People's Republic of China
1993–96
Senior UK Consultant, Test for English Majors (TEM) Validation Study, People's Republic of China
1995–98
Senior UK Consultant, Advanced English Reading Test (AERT), People's Republic of China
1999–04
Senior Consultant USAID Student Achievement Test Development Project, Egypt
2001–18
Consultant, Language Training and Testing Center (LTTC), Taiwan, including acting as international representative for the General English Proficiency Test (GEPT) research board

Learning and teaching

25 successful PhD completions
External examiner for 10 PhD theses

Measures of esteem

Joint Series Editor for Studies in Language Testing (SiLT) series, with Michael Milanovic (2013–14) and Nick Saville (2014–18)
Member of the Editorial Board for *Language Testing*
Academician of the Academy of Social Sciences
Member of the Steering Group of the English Profile Project, a major contributor to the revision of the Common European Framework of Reference for Languages (CEFR)
Chairman of the British Council Assessment Advisory Board

Awards

2013
Fellow of the Academy of Social Sciences
2014
Cambridge/ILTA Distinguished Achievement Award
June 2015
Officer of the Most Excellent Order of the British Empire (OBE) for services to English language assessment