Language Assessment Literacy and Competence Volume 2: Case Studies from Around the World

Also in this series:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

On Topic Validity in Speaking Tests
Nahal Khabbazbashi

Assessing Academic Listening: The Case of IELTS
John Field

Language Assessment Literacy and Competence Volume 2: Case Studies from Around the World

Edited by
Beverly Baker
University of Ottawa, Canada

and
Lynda Taylor
CRELLA, University of Bedfordshire

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Shaftesbury Road, Cambridge CB2 8EA, 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

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press & Assessment is a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781009504768

First published 2024

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

Printed in the United Kingdom by

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

ISBN 978-1-009-50476-8

Contents

Acknowledgements

It has been both a privilege and a pleasure for us to co-edit these two complementary volumes – Volumes 55 and 56 – focusing on *language assessment literacy and competence*.

We first discussed the idea of publishing an edited collection on this topic over afternoon tea in a Montreal café on a rainy day in October 2019. As researcher-practitioners based in the UK and North America we shared a strong interest in how theoretical and empirical research concerning language assessment literacy (LAL) was progressing at pace. We agreed that the time might be right to take stock of our current understanding of LAL for a range of stakeholders and consider how work in this area might need to develop further.

We are grateful to Dr Nick Saville, Director of Thought Leadership at Cambridge University Press & Assessment, for being open to the initial idea of a published volume on LAL within the Studies in Language Testing (SiLT) series, for which he was a co-editor. When the level of interest from potential contributors became clear, it was actually Nick's suggestion to explore producing two volumes instead of just one. This paired arrangement has worked inordinately well, allowing us to include a broad range of contributions from around the world. A key strength of the volumes lies in the fact that so many contexts are explored by those who know them best, leading to a rich offering of context-specific empirical research insights and reflections.

We express our gratitude to the more than 60 contributing authors who worked with us on successive drafts of their chapters over a 2-year period. Their readiness to engage with us through extensive email correspondence and occasional online Zoom/Microsoft Teams meetings knew no bounds, and their patience and diligence in working to revise their chapters in light of our editorial suggestions was humbling. We are also grateful to our authors for being so willing and enthusiastic to share something of their own voyages of LAL discovery in recent years. We are especially pleased to be able to include in both volumes a significant number of chapter abstracts supplied by the authors in languages other than English.

The two titles could not have reached publication without the dedication and skills of John Savage, Research Support Administrator – Publications, at Cambridge University Press & Assessment. Managing two SiLT volumes simultaneously through the complex editorial and print production processes

is no mean feat, and John deserves our deepest gratitude for keeping everything in the right place and on track.

Finally, we are grateful to the many colleagues worldwide – both past and present – who over the years have helped inform, shape and refine our own understanding and practice of LAL. We warmly commend the edited collections to the international language testing and assessment community, not least for the insights they might bring each of us for our own professional journeys in this field we share.

Beverly Baker and Lynda Taylor
January 2024

Series Editors' note

It is 30 years since the Studies in Language Testing (SiLT) series was first conceived, and in introducing Volumes 55 and 56 on language assessment literacy and competence (normally abbreviated to LAL), it is appropriate to reflect on one of the original objectives of the series: to extend the range and type of resources available to support LAL at a time when the concept was emerging and when language testing as an academic field was growing with an expanding community of professional practitioners.

This expansion was manifested by the establishment of the International Language Testing Association (ILTA) with its annual Language Testing Research Colloquium (LTRC), together with other professional bodies and associations that were constituted on a regional basis, such as Association of Language Testers in Europe (ALTE), Japan Language Testing Association (JLTA) and European Association for Language Testing and Assessment (EALTA) in the 1990s and early 2000s. By the end of the first decade of the 21st century there were already two world-class journals (*Language Testing* and *Language Assessment Quarterly*) and many prestigious Master's and PhD programmes on offer around the world.

Although LAL had its origins in broader educational contexts dating back to the early 1990s, the concept was enthusiastically taken up in language assessment during the 2000s, as part of the professionalisation of the field in general. The need for higher levels of theoretical knowledge and practical skills was strongly felt at a time when high-stakes language assessment was increasing in a wider range of societal contexts, including recruitment, healthcare, migration and citizenship. The SiLT series has made its own contribution in publishing 56 volumes offering both academic and practically oriented guidance and supporting the burgeoning interest in contemporary aspects of LAL.

This series was initiated by Dr Michael Milanovic in 1995 and he invited Professor Cyril Weir to become joint editor in 2003. In seeking authors for SiLT over the years, the Series Editors elicited topics to extend scope and coverage while maintaining the highest academic standards. Over the past 15 years the series has further broadened its reach, with contributions from more than 300 academics and practitioners from about 40 countries. In some cases, this has included topics that would not easily have been published elsewhere, including 12 PhDs, works on multilingual assessment, and international conference proceedings. The key aim of publishing high-quality PhD theses

has enabled emerging scholars to get their work into print and recognised; some have gone on to become senior leaders in their field. In short, the series has responded to the changing needs for LAL so it is fitting, therefore, that the last two volumes to be published in the SILT series should focus on contemporary aspects of language assessment literacy and competence as understood in the 2020s.

This is a timely contribution as the need for specialised knowledge and competences in language assessment has been expanding beyond educational contexts. This is reflected in the design of Volumes 55 and 56 covering *macro* and *micro contexts* and a broad range of multilingual assessment purposes and uses. Nowadays, the target audience for LAL is not only teachers and academics, but also policymakers and educational managers who oversee assessment regimes at a macro level and are responsible for delivering the intended outcomes and benefits for society.

In approaching this topic, Taylor and Baker undertook an extensive review of the field, seen through the lens of their own experiences as language testers who have grown up in the field as it has developed. In splitting their approach into the two separate volumes, they provide the reader with a detailed snapshot of the state of the art, illustrated with diverse personal and institutional perspectives, and case studies from a fascinating range of assessment contexts. This coverage illustrates the expanding geographical and institutional reach of LAL, and in total over 60 authors have contributed to these volumes, sharing insights from their own professional journeys. These insights and illustrations enable the reader to reflect on 'where we have arrived at' in our field and to consider where we might go next.

In Chapter 1 of Volume 55, Taylor outlines her own 'apprenticeship journey' in language assessment dating back to the late 1980s. In so doing, she sets the scene for the two volumes and their structure. As she explains, the aim is to provide relevant *research and reflections* on LAL *from the field* through the eight empirical research studies in Section 1 and the seven scholarly reflections in Section 2. Over the 35 years of her own journey, Taylor became increasingly aware that LAL is highly *context-based* and of the differing needs of the multiple stakeholders in assessment systems. She began reflecting on different ways to address this reality by adopting a 'community of practice approach' and this concept has informed the organisation of chapters and the insights from the research and reflections.

For Taylor, it became clear that technical knowledge is not enough, and that *collaborative engagement* is a key factor in developing the know-how and skills needed for an appropriate level of LAL for stakeholders – in their own contexts. Such collaboration is facilitated by paying adequate attention to factors such as *language and discourse* and ensuring *appropriacy for the context*. Stakeholder *beliefs and attitudes* need to be taken into account in building communities that are often interdependent.

These aspects are developed by Baker in Volume 56 through a compilation of 13 case studies to complement the 'more traditional' academic studies in Volume 55. A sharper focus on the voices of the stakeholders 'in context' is in keeping with the approach to LAL that the editors advocate, and a compelling argument for using case studies as a research methodology is made by Baker in her introduction.

We have two main aims in publishing these volumes: advancing our professional understanding of LAL and helping to open new avenues for future scholarship and practices in this area. This will be necessary to answer the question: what might the future hold for LAL in a rapidly changing world?

Since the original conception of this project in 2019, the world of language education has been buffeted by two global events that have made a lasting impact on society. The first was the COVID-19 pandemic in 2020 and the second was the arrival of Generative AI in the public domain, heralded by the open availability of ChatGPT from the end of 2022.

The lasting legacies of both events offer opportunities for innovation in language learning and assessment that need to be carefully considered and better understood if we are to take advantage of them. Of course, there are also attendant pitfalls and risks that have emerged and likewise need to be addressed to ensure that the innovations add value and have positive impacts. In other words, new aspects of language assessment are now emerging, and additional knowledge and competences will be added to the LAL repertoire in coming years.

Many commentators refer to the 'new normal' that was brought about by the pandemic. Not only was there a rapid uptake of existing educational technologies (EdTech) in 2020, but also the concept of 'hybridity' was crucial in the response to the closure of schools and offices. Hybrid working, hybrid classes and so on became part of the new normal, and this has accelerated innovations in EdTech, especially with deployment of automated systems using artificial intelligence (EdAI). In language assessment, for example, internet-based tests delivered to candidates in their own homes using *remote proctoring* was an important development during the pandemic.

Learning about and preparing to use AI effectively for assessment purposes has already been added to the LAL repertoire. Generative AI models using *large language models* (LLMs) and *deep neural networks* mean that there are both technical and ethical concerns that will require particular attention.

Other recent developments include a shift from large-scale standardised tests towards localised and personalised forms of assessment, with formative as well as summative functions, and with greater involvement of teachers in assessment processes. This suggests that *integrated learning and assessment* programmes will become more prevalent and language teachers will require additional knowledge and skills to deliver the intended benefits.

While the profession continues to focus on its traditional concern for *fairness*, the extension of the purposes and uses of assessments and the arrival of new technologies has provoked a greater concern for *social justice*, captured in the acronym JEDI: *justice, equity, diversity* and *inclusion*. These concerns come together, for example in the widespread uses of language tests for migration purposes and for obtaining citizenship, and in the exploration of new constructs related to multilingualism, such as plurilingual assessments.

These developments lead to another key question: in what ways should the language assessment community respond to the dilemmas and practical concerns that are emerging, and how can the profession continue to exert positive impacts? New forms of interdisciplinary collaboration, beyond current conceptualisations of LAL, will certainly play their part.

Nick Saville
Lynda Taylor

Cambridge 2024

Notes on contributors

Vivien Berry was Senior Language Assessment Researcher at British Council, London, when the research in her chapter in this volume was conducted. In addition to language assessment literacy, her research interests include oral language assessment, the subject of a forthcoming co-authored book. Her recent publications include articles and book chapters examining technology and oral assessment interviews.

Magnus Coney is an Assessment Manager in the Teaching Qualifications team at Cambridge University Press & Assessment. He has worked in the field of English Language Teaching since 2006 as a teacher, teacher trainer and examiner. He holds an MPhil in Research in Second Language Education, and his research interests include teacher education and the assessment of teaching.

Eddie Cowling is the Assessment Lead at the International Pathway College, University of York. His work primarily concerns assessment design, policy and procedure, with a focus on language testing. He is particularly interested in how collaboration and piloting can improve the quality of tests, and also has a keen interest in the role of technology in assessment. He is a Senior Fellow of the Higher Education Academy (HEA) and committee member of the BALEAP Testing Assessment and Feedback special interest group (TAFSIG).

Joaquín M Cruz Trapero is tenured at the Center for Higher Studies in Modern Languages of the Universidad de Jaén, Spain, where he has been in charge of language test design and validation, and foreign language teaching since 2011. His research focuses on the development and statistical analysis of rating scales. He is an affiliate of the Association of Language Testers in Europe (ALTE) and the European Association for Language Testing and Assessment (EALTA), and the director of the international postgraduate course Language Test Design and Validation at his university.

Margaret Early is an Associate Professor in the Department of Language and Literacy at the University of British Columbia. Her research focuses on multilingual and multimodal pedagogies across the curriculum, in multilingual school contexts (K-12). She co-edited the book *Identity Texts:*

The Collaborative Creation of Power in Multilingual Schools (Cummins and Early (Eds) 2011) and co-authored *Big Ideas for Expanding Minds: Teaching English Language Learners across the Curriculum* (Cummins and Early 2015).

Marta García is Associate Professor in Cursos Internacionales at the Universidad de Salamanca, where she teaches Spanish and is also responsible for the examination LanguageCert USAL esPro, a test of general Spanish language for professional purposes. She is a member of the standing committee of the Association of Language Testers in Europe (ALTE) and has recently taken over the coordination of the Language for Specific Purposes Special Interest Group (LSP SIG) of ALTE. She has worked for many years in the language testing industry, particularly for Spanish as a foreign language. She is also Adjunct Professor in the Modern Languages Department at the Universidad de Salamanca.

Frank Giraldo is a professor at the Foreign Languages Department of Universidad de Caldas in Manizales, Colombia. He teaches academic writing and language assessment at the undergraduate and graduate levels. He has published in the area of language assessment literacy, particularly in the Colombian context. His main research interest is the interface between language assessment literacy and teachers' professional development.

Mojtaba Heydari has a PhD in Applied Linguistics. He has several years of experience in the field of testing and teaching languages. His research interests include second language assessment (SLA) and the use of new technologies in SLA. Recently, he has been working on the application of data science in SLA. His publications include an article in the journal *Assessing Writing* and three edited book chapters published by Routledge and Palgrave Macmillan.

Connie Lam James is a recent graduate of the Languages, Cultures and Literacies program at Simon Fraser University, Canada. She is currently the Administrative Manager, Undergraduate Programs, College of Education and Social Transformation, J.F. Oberlin University (Tokyo, Japan), Vancouver Office, Canada. Her research and teaching interests include working with international students, multilingual students, and students in academic transition.

Jiyoon Lee is an associate professor in the TESOL program at the University of Maryland, Baltimore County (UMBC) in the Department of Education. She specialises in the fields of language education and teacher preparation, with a focus on language assessment and language teacher development. Her research focuses on language assessment literacy, exploring its impact on both teachers and learners in language teacher education.

Kathryn Macfarlane completed her PhD in Applied Linguistics at Monash University, Australia. She is also the founder of Reading's a Breeze! Ltd (www.readingsabreeze.com), a publishing company dedicated to producing multilingual read-aloud eBooks for young learners of additional languages. As a language consultant, she has designed and delivered professional learning for language teachers in Australia and New Zealand since 2010. Her research interests include language teacher development, second language acquisition in instructed contexts, and parenting strategies for raising children bilingually.

Fahimeh Marefat is a professor of Applied Linguistics at Allameh Tabataba'i University, Tehran, Iran. Her area of interest includes assessing writing, genre studies, and systematic reviews. She has contributed to some edited volumes published by Routledge, Palgrave Macmillan and Springer. A recent article appeared in Studies in Second Language Learning and Teaching.

Waldemar Martyniuk is Professor at the Institute of Polish Language and Culture for Foreigners of the Jagiellonian University in Kraków, Poland. He holds a PhD in Applied Linguistics and is a teacher trainer, author of several textbooks, curricula, and testing materials for Polish as a foreign language. He has been a visiting professor and lecturer at several universities in Germany, Switzerland (Basel), and in the US (Stanford University). He has been seconded to the Council of Europe, Language Policy Division (2005–06), and as Executive Director of the European Centre for Modern Languages (ECML) in Graz, Austria (2008–13). He is Chair of the Board of Trustees at the Association of Language Testers in Europe (ALTE).

Ben Naismith is an Assessment Scientist on the Validity, Efficacy, and Outreach team of the Duolingo English Test. He has worked extensively in the field of English language teaching in numerous contexts as a teacher, teacher trainer, materials developer, assessment specialist, and researcher. He holds a PhD in Applied Linguistics, and his research interests include instructed second language acquisition, learner corpus research, language assessment, teacher education, and lexical development.

Rob Playfair is an English for Academic Purposes (EAP) Practitioner and PhD researcher at Birkbeck College, University of London. He is from London and taught English as a Foreign Language (EFL) for over a decade in East Asia, South America and Europe before returning home. His research interests include assessment impact and ethnography, and for his PhD he is exploring the role of EAP teacher knowledge in shaping and being shaped by practice. He was a founding member of the BALEAP Testing Assessment and Feedback special interest group (TAFSIG).

Lorenzo Rocca is responsible for the language projects department at the Società Dante Alighieri in Rome. He worked for many years at the University for Foreigners of Perugia, engaging in test development and teacher training. His research interest is focused on the migration context; with regard to this focus he has been the Chair of the Association of Language Testers in Europe Language Assessment for Migration and Integration Special Interest Group (ALTE LAMI SIG) since 2008 and a member of the Council of Europe Linguistic Integration of Adult Migrants (LIAM) project since 2014. In 2020 he was awarded the national scientific qualification of Associate Professor.

Graham Seed is the Secretariat Manager of the Association of Language Testers in Europe (ALTE), and Senior Research Manager at Cambridge University Press & Assessment. He has also worked as Senior Assessment Manager for various test products and in language education in different European countries. He has a Master's in Language Testing from Lancaster University, and current research interests include plurilingualism in language assessment and the operationalisation of the Common European Framework of Reference for Languages (CEFR) in language assessment.

Susan Sheehan is Senior Lecturer in TESOL at the University of Huddersfield. She teaches at both the graduate and undergraduate levels and is the Course Leader for the BA TESOL degree. Susan organised the first symposium presented by the International Language Testing Association's Language Assessment Literacy Special Interest Group. The symposium was entitled: 'Towards a model of language assessment literacy'. Her research interests include language assessment literacy, classroom-based assessment and teacher cognition.

Richard Spiby has worked at the British Council in London with the Assessment Research Group since 2016. He works on the development and validation of British Council tests. His main responsibilities include overseeing operational analysis and investigating the receptive skills components for new and existing tests. He also works on a variety of assessment development and training projects worldwide. His areas of interest include language assessment literacy, test-taker strategies, test analysis and inclusivity in language assessment.

Valia Spiliotopoulos is an Assistant Professor and Director of the Post-Secondary Immersion Research Group at the Official Languages and Bilingualism Institute (OLBI) at the University of Ottawa, Canada. Her research focuses on content and language integrated learning/immersion

at the university level, teacher education, and assessment in bi/multilingual contexts. Her publications have appeared in *Higher Education Research and Development, TESL Canada Journal,* and *The Canadian Modern Language Review*.

Olga Ukrayinska is Associate Professor at Kharkiv Skovoroda National Pedagogical University, Ukraine, where she teaches Introduction to Language Testing and Methods of Foreign Language (FL) Teaching at the graduate and undergraduate levels at the Faculty of Foreign Philology. She holds a PhD in FL Teaching and is an expert and item writer for the Independent School-Leaving Examination in the Ukrainian Center of Evaluation of Quality of Education. Her research interests include FL teaching and assessment in tertiary education, specifically in developing student teachers' language assessment competency.

Koen Van Gorp is an Assistant Professor in the Department of Linguistics, Languages, & Cultures and Head of Research in the National Less Commonly Taught Languages Resource Center (National LCTL Resource Center; NLRC) at Michigan State University. He is the founding co-editor of *TASK: Journal on Task-Based Language Teaching and Learning* and Treasurer of the International Association for Task-Based Language Teaching. His research interests include task-based language teaching and assessment and multilingual education.

Saskia Van Viegen is Associate Professor and English as a Second Language (ESL) Coordinator in the Department of Languages, Literatures and Linguistics at York University in Toronto, Canada. Her research engages with bi/multilingualism in education, language assessment, and language teaching and learning in post-secondary contexts. She co-edited the book *Plurilingual Pedagogies: Critical and Creative Undertakings for Equitable Language (in) Education* (Lau and Van Viegen (Eds) 2020) and edits the journal *Critical Inquiry in Language Studies*.

Dina Vîlcu teaches and assesses Romanian as a foreign language (RFL) at Babeş-Bolyai University, Cluj-Napoca. She teaches general language and Language for Specific Academic Purposes (LSAP) courses (for social and humanistic sciences and arts), as well as tailor-made courses for interpreters and translators. She also teaches courses on Language for Specific Purposes (LSP) and language assessment at MA level. She is interested in the challenges involved in language assessment for less widely tested languages (LWTL), currently chairing the Association of Language Testers in Europe (ALTE) LWTL SIG.

Carolyn Westbrook is a Test Development Researcher at the British Council. She has worked on a number of testing and assessment development and validation projects, and has provided language assessment literacy training for teachers around the world. She has also worked as a teacher, teacher trainer and materials writer, and has a wide range of experience in General English, Business English, English for Academic Purposes (EAP) and English for Specific Purposes (ESP). Her research interests are integrated assessment as well as EAP and ESP assessment.

1 Introduction

Beverly Baker
University of Ottawa, Canada

Lynda Taylor and I are proud to introduce Studies in Language Testing Volume 56, *Language Assessment Literacy and Competence: Case Studies from Around the World*. This volume is a companion to Volume 55, *Research and Reflections from the Field*. Together the two volumes focus on current understandings and practice in the development of language assessment literacy (LAL) and competence as reflected in a series of empirical research papers, case study reports and individual reflections.

Here I describe how this volume came about, sharing our understandings of case studies together with our motivation in presenting a collection of cases to complement the traditional chapters of Volume 55. A part of this motivation includes redirecting attention to the centrality of *context* in our field, something that case studies are especially effective in highlighting and exploring. With this volume we aimed to dispel narrow perceptions of the usefulness of case studies, drawing the reader's attention to some of the key contributions of this particular collection of cases to scholarship in language assessment literacy and competence.

In our planning we did not originally set out to dedicate a volume to case studies *per se*. In mid-2021, while reading through the dozens of submissions for our proposed volume, Lynda and I were struck by the diversity represented in the abstracts in terms of geography, data sources, collaborators, and approaches to investigating language assessment knowledge building. We saw the potential to profile some of these scholars and their work in a new way. As a result, we contacted a selection of these scholars to ask if they would consider developing their submission not as a traditional book chapter but as a report of their particular case – shorter but with more concentrated details of their specific context than they might normally report. While many of our authors did not necessarily set out to do case study research, this does not need to be done *a priori*: as Yin (2013) reminds us, cases can be conceptualised as a way to describe situated phenomena. Our contributors were willing to reconceptualise their work in this way, providing the particularities of their situation and endeavoring to share what might resonate for others in their own circumstances across our international community. We commend them for taking up this challenge. Their successful efforts here can encourage others to feel confident about doing something similar in their own contexts.

Defining and operationalising case studies

In the words of Gerring, 'Regretfully, the term "case study" is a definitional morass' (2004:342), suggesting that it is so amorphous a concept as to defy description. Despite this, he provides what we find to be a very useful definition, straightforward yet broad in scope: 'A case study is best defined as an in-depth study of a single unit (a relatively bounded phenomenon) where the scholar's aim is to elucidate features of a larger class of similar phenomena' (2004:341).

This idea of relative as opposed to absolute boundaries allows for remarkable conceptual flexibility, and this volume effectively demonstrates the various forms a case can take, such as the following:

- the case of *a single school* that implemented an innovative approach to FL learning in Australia, where the teachers and students are learning the language at the same time (**Kathryn Macfarlane**, Chapter 6)
- *an organisation* (ALTE) with a mission of disseminating knowledge and expertise in language assessment to a wide variety of stakeholders, through the provision of resources, educational initiatives and policy collaboration and advocacy (**Graham Seed**, **Waldemar Martyniuk** and **Lorenzo Rocca**, Chapter 10)
- *an activity* to design analytic scales for teachers in various schools within a public language school network in the Canary Islands (**Joaquín M Cruz Trapero**, Chapter 11)
- *a group of three teacher trainers* on the Cambridge CELTA (a teaching certification course), using an innovative 'lesson study' approach for their professional development (**Magnus Coney** and **Ben Naismith**, Chapter 12).

Many of the cases reported in the volume concern groups of teachers, at various levels and stages of their careers, spanning the globe. Such cases represent *local testing and assessment* defined by Dimova, Yan and Ginther as 'one whose development is designed to represent the values and priorities within a local instructional program and designed to address problems that emerge out of a need within the local context in which the test will be used' (2020:1). For example, **Frank Giraldo** (Chapter 7) investigated a group of practising high school teachers in one area of Colombia, tailoring training to the teachers' specific needs and goals. **Valia Spiliotopoulos**, **Saskia Van Viegen**, **Margaret Early** and **Connie Lam James** (Chapter 2) share the activities of a university–school district professional learning community working on the implementation of new English language learning standards in British Columbia, Canada. **Olga Ukrayinska** (Chapter 3) explored the development of self-assessment skills of pre-service English and French teachers in Ukraine. We applaud her for succeeding in contributing to this

volume despite a number of practical obstacles. We appreciate her continued dedication and that of her colleagues to teacher education in the face of ongoing conflict.

The cases presented here are not all delineated by geography. The projects described by **Carolyn Westbrook** and **Richard Spiby** (Chapter 9) as well as by **Seed et al** (Chapter 10) and **Rob Playfair** and **Eddie Cowling** (Chapter 14) do not focus on a specific geographical context. Instead, they report initiatives designed for groups with a specific common need. Contexts here can also be described in terms of 'micro-contexts at the level of the individual or macro-contexts of the wider national, political, and educational system in which that individual is located' (Lai and Fjørtoft 2022:1). In our volume, **Jiyoon Lee** (Chapter 5) investigates the reflections of pre-service teachers with a classroom-based project in the United States (a micro context) while **Dina Vîlcu**, **Koen Van Gorp** and **Marta García** (Chapter 13) explore the revision of two English for Specific Purposes tests in Romania and Spain with reference to an international guidelines document (perhaps best described as a meso context). Regarding large-scale, commercially available standardised tests (macro contexts), **Mojtaba Heydari** and **Fahimeh Marefat** (Chapter 4) employed an innovative data mining procedure to explore comments from a social networking discussion group of test-takers in order to uncover their (mis)conceptions of IELTS as well as their concerns with the test taking experience – as expressed to their peers rather than to researchers.

The importance of case studies in advancing knowledge

The definitions of cases above refer not just to how cases can be delineated, but to their uses and their usefulness in contributing to knowledge. Stake (1995) defines cases as intrinsic (studying a unique instance), instrumental (indicative of larger phenomena) and collective (used as a collection to paint a portrait of an issue). We see these categories as complementary rather than mutually exclusive. Firstly, the *intrinsic* nature of cases means priority is given to the insights that can be gained from studying a relatively limited object or grouping *in situ*. Flyvbjerg writes: 'The advantage of a case study is that it can "close in" on real-life situations and test views directly in relation to phenomena as they unfold in practice' (2006:235). Crowe et al echo this sentiment, calling case study research '… an in-depth appreciation of an issue, event or phenomenon of interest, in its natural real-life context' (2011:1).

Cases are *instrumental* in that they 'illustrate broader lessons that may be learnt' (Crowe et al 2011:1), and *collective*, in the sense that knowledge of collections of cases can be viewed as the basis for the development of expertise. This is similar to what Yin (2013) terms 'analytic generalization',

a process of abstracting from a case outward. Flyvbjerg argues that '… a scientific discipline without a large number of thoroughly executed case studies is a discipline without systematic production of exemplars, and a discipline without exemplars is an ineffective one' (2006:219). Stating that the case study 'may be central to scientific development' (2006:228), he provides examples of research that has advanced humankind through individual experiments and case studies as opposed to formal generalisation and hypothetical-deductive models: 'That knowledge cannot be formally generalized does not mean it cannot enter into the collective process of knowledge accumulation in a given field or in a society. A purely descriptive, phenomenological case study … can certainly be of value in this process and has often helped to cut a path toward scientific innovation' (2006:227). He takes the example of the law of gravity, which was established not through controlled experimentation with random samples, but through the selection of the two cases of lead and a feather to determine whether they would fall with equal velocity.

As can be inferred above, one's understanding of the term 'generalisable' depends on one's epistemological positioning as critical, interpretive or positivist (Crowe et al 2011). Flyvbjerg (2006) describes a positivist orientation as one which privileges 'context-independent' knowledge. An extreme view of this would be the concept of an ideographic universe, described by Gerring where 'absolutely nothing can be learned about one unit by studying another' (2004:351). Those taking critical and interpretivist positions tend to view knowledge as essentially indivisible from the context in which it is generated, a view that is consistent with the current post-positivist orientation of our field and in the social sciences more generally. Judging from the recent proliferation of context-bound empirical studies in language assessment, we make two observations: i) that scholars in our field are taking a generally interpretative approach, acknowledging the inseparability of context from their discoveries, and ii) that despite this, they do see the instrumental and collective value in sharing these context-bound experiences with others.

In other words, we see evidence from the very existence of this work that scholars across our field adhere to the belief that 'a context-specific problem … can be solved while advancing methodological and theoretical knowledge in a range of domains …' (Lai and Fjørtoft 2022:2). Of course, discoveries in one context may require a certain amount of imagination to apply to other settings. Lai and Fjørtoft describe the concept of 'boundary crossing' from educational theory as the highly productive movement of innovations among contexts: 'ideas or practices arising in one educational context are therefore not seamlessly integrated into others; instead, they are subject to processes of reinterpretation, hybridization, or transformation' (2022:2).

This discussion of the value and potential of cases brings us to a further observation regarding an unfortunate convention we still see regularly

in our field: after scholars have presented their in-depth case study at a conference or in a publication, and have argued, often convincingly, for the insights their discoveries can offer their peers and the knowledge base of their field, they then finish by stating that their case study approach may be considered a 'limitation' of their research – that because it is a case, it 'cannot be generalized'. For us, this type of statement amounts to conducting a study through a critical or interpretive framework and then veering into positivism at the last minute, at the very moment the researcher is arguing for the value of their work! The intrinsic, instrumental and collective benefits of case studies which motivated the study in the first place are discounted with reference to some sort of context-free standard. Such statements risk perpetuating the idea that the gold standard in research of any kind is the randomised controlled experimental study, and that knowledge is only generated by conceiving of a case as a representative sample which can represent the population as a whole. Can we rid ourselves of this type of self-defeating and incoherent statement? Flyvbjerg (2006) challenges this misunderstanding of generalisation and its role in knowledge production, contending the following:

> Social science has not succeeded in producing general, context-independent theory and, thus, has in the final instance nothing else to offer than concrete context-dependent knowledge … Predictive theories and universals cannot be found in the study of human affairs. Concrete, context-dependent knowledge is, therefore, more valuable than the vain search for predictive theories and universals … formal generalization is overvalued as a source of scientific development, whereas the "force of example" is underestimated (223–228).

He evokes Beveridge, who contends that there 'are more discoveries stemming from the type of intense observation made possible by the case study than from statistics applied to large groups' (1951:236).

Insights gained from this collection

In summary, this collection demonstrates strong motivation to learn more about language assessment, at least for the diverse actors involved in language education. Desire to learn resonates through multiple contributions here. In terms of the construct, we see the term 'literacy' still used most often, with other terms such as 'competence' occasionally applied. Regardless of the terms used, it does not appear to be a controversial idea that different stakeholders' knowledge can be described in terms of differing but complementary profiles. We also observe a movement from questions related to the constituent elements of LAL – the WHAT (e.g., Kremmel and Harding 2020) – to questions of HOW to facilitate this development.

We see an awareness that there is no 'one size fits all approach' to LAL development for all stakeholders. In this volume, LAL development is addressed with a nascent professional learning community (**Spiliotopoulos et al**, Chapter 2), through Learning Study (**Coney** and **Naismith**, Chapter 12), and with online interactive workshops (**Giraldo**, Chapter 7), to name just some examples. We also see examples of development through collaboration, such as between teachers and their learners (**Lee**, Chapter 5, **Macfarlane**, Chapter 6) and among testing experts, teachers, and subject matter experts in the development of a specific purposes medical language assessment (**Vîlcu et al**, Chapter 13). In addition, the voices of test-takers themselves are recognised as sources of learning for language test developers themselves (seen in **Heydari** and **Marefat**, Chapter 4, among others).

Self-reports seem to be the preferred investigative tool represented here. For example, in their efforts to build an international piloting network amongst English for Academic Purposes practitioners, **Playfair** and **Cowling** (Chapter 14) used interviews to learn about the needs and challenges of their colleagues. In addition to self-reports, we see a place for an increase of observational studies including observation of formative assessment practices in the classroom (such as **Susan Sheehan** and **Vivien Berry**, Chapter 8, and **Ukrayinska**, Chapter 3) and collection of assessment-related products (see **Cruz Trapero**'s work in Chapter 11, where rating scales created by teachers serve as proof of learning).

These represent only a few of our own insights about the collection as a whole and what it represents for us and our own understandings, as well as for the field of language assessment more broadly. However, in the interests of 'boundary crossing', we invite our readers to interpret and draw their own insights from this collection according to their individual needs and interests, perhaps even following up with our authors for any details that might be helpful for a productive reimagining of this work in other contexts. Above all, we believe this collection represents a recognition and affirmation of the value of cases in our field, both as a scholarly language and as a way of enabling our progress as a collective.

References

Beveridge, W I B (1951) *The Art of Scientific Investigation,* London: Heinemann.

Crowe, S, Cresswell, K, Robertson, A, Huby, G, Avery, A and Sheikh, A (2011) The case study approach, *BMC Medical Research Methodology* 11, Article 100.

Dimova, S, Yan, X and Ginther, A (2020) *Local Language Testing: Design, Implementation, and Development,* London: Routledge.

Flyvbjerg, B (2006) Five misunderstandings about case-study research, *Qualitative Inquiry* 12 (2), 219–245.

Flyvbjerg, B (2021, May) Yes, you can generalize from a case study, despite what your supervisor says, *Geek Culture*, available online: medium.com/geekculture/yes-you-can-generalize-from-a-case-study-f8000bd647b9

Gerring, J (2004) What is a case study and what is it good for?, *The American Political Science Review* 98 (2), 341–354.

Kremmel, B and Harding, L (2020) Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the language assessment literacy survey, *Language Assessment Quarterly* 17 (1), 100–120.

Lai, M K and Fjørtoft, H (2022) Problematising and understanding contexts in educational research: An international perspective, *Studies in Educational Evaluation* 74, 1–4.

Stake, R (1995) *The Art of Case Study Research*, London: Sage Publications.

Yin, R (2013) *Case Study Research: Design and Methods* (Fifth edition), Thousand Oaks: Sage Publications.

2 Teacher professional learning communities and language assessment reform in western Canada

Valia Spiliotopoulos
University of Ottawa, Canada

Saskia Van Viegen
York University, UK

Margaret Early
University of British Columbia, Canada

Connie Lam James
Simon Fraser University, Canada

This case study reports on a university–school district collaborative inquiry on the implementation of a new English language proficiency assessment framework for K-12 education in British Columbia (BC), Canada – the BC *English Language Learning (ELL) Standards* (BC Ministry of Education 2017). This case study focuses on:

- Insights from a year-long study of teachers' field testing and professional learning of the BC ELL Standards in four school districts
- Activities and features of effective professional learning to build educators' language assessment literacy for supporting ELLs in BC education
- Mutual benefits of knowledge-sharing and knowledge creation in university–school district partnership projects and the importance of developing professional learning communities to advance teachers' language assessment literacy
- Key recommendations regarding initial and ongoing assessment, such as the use of exemplars, alignment with the provincial curriculum, and the inclusion of current inclusive, multilingual perspectives on language teaching and learning

Introduction

Immigration, internationalization, and increased student mobility have expanded the sociolinguistic landscape of British Columbia (BC) education, bringing greater attention to teachers' and students' language teaching and learning needs. Whereas researchers in applied linguistics and education have documented the contributions of developing teachers' language assessment literacy (LAL) to student learning and achievement (Coombe, Troudi and Al-Hamly 2012, Coombe, Vafadar and Mohebbi 2020, Vogt and Tsagari 2014, Volante and Fazio 2007), professional learning needs persist (Garnett 2008). In BC, 68,982 students were identified by schools as English Language Learners (ELL) in 2019, comprising 13% of the student population (BC Ministry of Education 2019). This superdiverse educational context (Li, Anderson, Hare and McTavish 2021) highlights the need for professional knowledge and expertise in language assessment, to shape teachers' knowledge and practice of language teaching in BC classrooms.

To address this need, in 2017 the provincial Ministry of Education launched the BC *English Language Learning (ELL) Standards* (BC Ministry of Education 2017), based on earlier draft standards developed at the district level, to be used by ELL specialist teachers across all school districts to guide language teaching and assessment. Drawing on insights gathered during a university–school district partnership to provide professional learning and to field test these ELL Standards, we report insights on teachers' developing language assessment literacy and needs and concerns.

Language assessment in BC education

Following an assessment for learning model, the ELL Standards (2017) were designed to guide classroom-based language assessment (CBLA) of students whose primary or home language is other than English, and who need 'explicit and specific language support to access the curriculum of content-areas (such as English Language Arts, Social Studies, Mathematics, Science, etc.) at their grade level' (BC Ministry of Education 2017:5). The standards (2017:5) are comprised of English language proficiency (ELP) descriptor scales that articulate developmental stages of language learning across three domains of language use (reading, writing, oral language). Each domain uses a five-level scale, reflecting a continuum of development from the 'beginning' (level 1) to 'bridging' (level 5), which represents the level of language proficiency of peers of the same age and grade level. Separate scales exist for primary, intermediate, and secondary levels.

The scales are designed for use by teachers to conduct both initial and ongoing assessment. The Ministry (2017:4) articulates that the purposes of the ELL Standards are to:

- complement the various approaches school districts have developed for planning and carrying out ongoing ELL student support, including assessment and classroom adaptations
- provide a common language for describing language proficiency among schools and school districts for clarity and consistency throughout the province
- facilitate communication with ELL students and their parents
- encourage collaboration among all educators regarding the ELL students' language proficiency and support needs.

Effective implementation and use of the ELL Standards require teachers' language assessment literacy[1], or what Inbar-Lourie (2008:389) determines as 'the capacity to ask and answer critical questions about the purpose for assessment, about the fitness of the tool being used, about testing conditions, about what is going to happen on the basis of the results'. Teacher awareness of the social consequences and impact of their assessment practice and related decisions is critical (Scarino 2013, Stille, Jang and Wagner 2015) because teachers' interpretations and decisions about students' language abilities have impact both in the classroom and beyond (Edelenbos and Kubanek-German 2004, Jang et al 2015). While CBLA promotes teachers' agency in assessment in general, it recognizes a need to develop teachers' assessment literacy, increasing teachers' knowledge of and communication about language proficiency development (Cummins et al 2009, Jones and Saville 2009) and their professional judgements about learners' language development needs (Cumming 2009, Davison and Leung 2009). More specific to the BC context, the lack of a required course on ELL education and language assessment in teacher education programs, as well as the variability across districts in initial and continuing language assessment practices, suggest the need for guiding principles and a common language around language assessment.

Project design and activities

Following the launch of the ELL Standards in 2017, ELL Coordinators from four BC school districts partnered with university-based researchers to engage in professional learning and inquiry concerning language assessment practice at both the elementary and secondary levels. The collaboration was informed by key principles in community-engaged research (Tremblay and Hall 2014), to build research-practice partnerships and develop situated, context-specific knowledge and understanding.

1 Literacy is used synonymously with the term competence throughout the manuscript.

Thirty-five ELL teachers joined the project, incorporating the standards into their practice and documenting their language assessment activities for one school year as part of the CBLA field testing process. Participating teachers selected marker students[2] and created a portfolio with evidence of students' linguistic performances that they assessed using the ELL Standards. They attended four cross-district professional learning sessions, comprising a half-day of professional learning with other participating teachers, co-facilitated by university-based researchers and district ELL Coordinators/Consultants.

Each professional learning session was guided by an explicit focus. The first session involved a broad overview of the ELL Standards and focused on assessing writing. The second and third sessions focused on assessing reading and oral communication, respectively, and the final session engaged participating teachers in sharing overall reflections and holistic feedback on the use of the ELL Standards. For each session, teachers were placed into small groups according to grade level and/or district, which were later mixed for cross-level and -district sharing. Using artifacts of student work, teachers engaged in moderated marking and benchmarking activities using the ELL Standards to make explicit their tacit professional knowledge (Rust, O'Donovan and Price 2010).

Data collection and analysis

To better understand the potential for development of teachers' LAL, the university-based researchers gathered data using observations and fieldnotes during the four professional learning sessions. At the end of the study, teachers completed an online survey (n = 35) about perceptions of the standards. For example, one of the questions in the survey asked what the standards could potentially address or accomplish for BC teachers. Additionally, select teachers participated in open-ended interviews (n = 7) for in-depth discussion of their perceptions. Teachers also shared their assessment portfolios with the research team. Survey and interview data were analysed to generate themes within the data (Saldaña 2015), and triangulated with insights from the observations and fieldnotes gathered during the professional learning sessions. Below, we present insights from the report that illustrate teachers' developing LAL and teacher perceptions of the ELL Standards.

2 Marker students are the students who were selected by the teachers to participate in this study. Their consent enabled the teachers to gather assessment artifacts, which were compiled into a portfolio.

Results

From what teachers expressed during the sessions, there is some evidence that they perceived the sessions to be a potential space for further development of assessment literacy. The excerpts from the surveys and interviews below illustrate participating teachers' engagement with critical questions about the purpose, validity, context, and consequences of using the standards. By sharing experiences and insights using a common language, teachers articulated their professional judgements concerning assessment and collaborated with colleagues to examine assessment issues. The following sections highlight key themes revealing teachers' reports of their growth.

Distinguishing between diagnostic, formative, and summative assessment

Participating teachers applied and analysed the ELL Standards through guided practice during the professional learning sessions. These sessions served as a place where teachers could represent and demonstrate their knowledge and assessment competence. Teachers described their use of the standards for diagnostic purposes, illustrated by the following statement from an ELL Helping teacher on the survey: 'Results are used for placement purposes and sent to schools to assist teachers in creating appropriate programs and types of support for teachers.' Although many teachers were already familiar with the use of standards for diagnostic purposes before engaging in these sessions, teachers reported that these professional activities helped them learn how to use the standards for formative assessment purposes. As an elementary-level (1–7) ELL teacher shared through the survey: 'Assessments are used to let teachers know what the student is lacking, therefore, lessons and practices will be centred on the student's needs.'

Participating teachers further explained the use of the ELL assessment standards for summative purposes, and as a way to bridge summative assessment from one year with diagnostic assessment for the following year:

> So what I would do this year is assess all the students at the end of let's say Grade 3 or Grade 4 … Then I take that information and that's what their ELL level is for next year.
>
> (Interview; elementary-level (K-5) ELL teacher)

Regarding integrating language assessment into their practice, participating teachers noted a lack of consistency across assessment and evaluation procedures between the ELL Standards and the curricular competencies, particularly in subject content areas.

Assessment literacy through a shared assessment language

Participating teachers recognized the potential for individual and collective capacity-building to meet students' needs by learning about language assessment. As one teacher shared:

> I am very excited that at least the teachers in BC will have a shared language to discuss student growth and achievement. It will never be the only assessment tool (particularly not in its current state!) but it will be incredibly useful to turn to, particularly if there are student exemplars at each grade and ELL level to consult as a reference.
>
> (Survey; secondary ELL teacher)

Participating teachers also thought that the ELL Standards can help teachers who are supporting students integrated into mainstream classes, and can facilitate communication through a shared language between ELL specialists and classroom teachers. The following comment illustrates this:

> It's also been very useful in terms of communicating to other teachers, like non-ELL teachers, but classroom teachers … 'cause a lot of classroom teachers don't really know necessarily what we're focusing on explicitly.
>
> (Interview; elementary-level (K-7) Resource & ELL teacher)

Assessment literacy through reflective practice

By gathering samples of student work, sharing student portfolios with colleagues, and articulating the rationale for their professional judgements, teachers were engaging in both individual and communal reflective practices. One teacher described how the documentation process was effective in this regard:

> As a result of using this documentation tool, my instructional practices have changed to include a perspective on general topics, then specific examples within these topics (e.g., when participating in outdoor learning (science).
>
> (Survey; elementary-level (kindergarten) Resource teacher)

Teachers gained understanding of the need for guidelines and support for engaging in language assessment. This is expressed by one educator regarding reading:

> How do we measure reading, comprehension? What kind of post-reading tasks do we set […]? Should it be an interview to the teacher … you know to

> test reading comprehension? Should it be a writing task, and what is it really we're measuring then? So there can be a lot of confusion for educators.
>
> (Interview; secondary-level ELL teacher)

Importantly, teachers critically reflected on the fairness and reliability of their assessment practices when using the matrices and analysing their grading practices both individually and collectively. Questions arose about whether the English monolingual standards normalize native speakerism, and erase bilingualism since translanguaging practices are not recognized. The standards do not represent the sociolinguistic ecology of current classrooms and schools, and they can be used for gatekeeping processes. Some questioned whether the performance levels may be somewhat arbitrary, or based on native-speaker norms:

> I am not happy with the summative assessments for levelling and placing students. I don't think they are culturally inclusive. I think they "label" rather than provide diagnostic information for teachers. Parents and students see them as a "pass/fail" "test/exam" and misunderstand what ELL support is.
>
> (Survey; secondary-level ELL teacher)

Another secondary-level ELL and Academic Writing teacher noted in an interview that 'native speakers don't have that, native speakers don't always support a thesis sentence. Like give me a break. They repeat themselves and they get off topic and their writing is all over the place.'

In sum, teachers not only engaged in reflective practice to motivate their own ongoing development in assessment literacy; teachers also engaged in critical reflection on the inconsistencies between educational norms and students' language practice. The use of the standards – and the implications and consequences of their use – encouraged teachers to critically reflect on how ELL students are perceived by teachers, tests, and schools according to institutional categories and labels (Gunderson 2021, Talmy 2015).

Assessment literacy development through collaboration

A consistent theme that emerged from the feedback after each professional learning session, as well as in the interviews and surveys, was that the collaboration with other educators was a key professional learning and community-building activity that helped develop their literacy.

> Just hearing about what other people are doing. You know often we just get kind of insulated in our school or our classrooms. […] it just makes us

> stop and think about […] how can we all work together to make learning better for all our students.
>
> (Interview; elementary-level (K-5) ELL teacher)

The benefits of this collaboration across schools, districts, and grade levels were perceived as so effective in impacting student learning that teachers expressed an interest in continuing with this type of professional learning activity:

> I'm just wondering how we can continue to do those sorts of things, building community, not just within schools but districts and then even within the province across different districts. This has been, it's been really amazing, and you know having the opportunity to speak to speak to colleagues out, out of, out of your own school.
>
> (Interview; secondary-level ELL teacher)

As seen above, this sense of community-building with the same group of teachers, albeit from diverse contexts, was reinforced because all the teachers had a common goal. As a secondary-level teacher shared in an interview, she 'grew pedagogically by hearing other people's perspectives'. Participating teachers expressed that the group feedback and discussions on assessment tasks, descriptors, and professional judgements not only helped validate and change their instructional practices, but enabled them to go back to their own contexts and share with their immediate colleagues, thereby strengthening the professional collaborations and learnings within their own school community. The emergent professional learning community (DuFour and Eaker 1998) helped generate positive dialogue among educators and a shared understanding about language proficiency development and assessment:

> I have a job share partner and my coworkers as well, and I think it's, for me it was important to share these descriptors with them just to kind of help just direct them and then direct me as well, just to help each other.
>
> (Interview; elementary-level (kindergarten) Resource & ELL teacher)

An additional theme that emerged from the various sources of data was that the collaboration activities with the university partners helped participating teachers engage in practices informed by research, which further supported their perceived professional growth in ELL assessment:

> So it's been rich, and I also like the fact that there were a couple times where we were given an opportunity to read an article, some, some type of research that's been done related to language and learning and respond to it.
>
> (Interview; elementary-level (K-5) ELL teacher)

The benefits of meaningful collaboration with other professionals in their school and beyond is consistent with theory and research in professional learning communities (PLCs) as key sources of professional development for teachers (DuFour and Eaker 1998). Not only did the educators appreciate new opportunities for learning, but they were also appreciative of having their voices and insights heard by others who shared similar professional values and vision. As previously mentioned, the approach used in this university–community partnership was not a 'sit and get' approach, but one that encouraged collaborative inquiry, where university colleagues had as much to learn from educators in the field:

> I'd really like to know how we can continue doing the types of things we've been doing throughout the workshop and having the you know teachers and educators, consultants and academics in the same room, not just academics from one institution but various institutions has been really, really transformative, really supportive.
>
> (Interview; secondary-level ELL teacher)

Conclusions and recommendations

The important findings of the field testing and professional learning of the revised ELL Standards have helped administrators, practitioners, and researchers carefully consider next steps in assessment reform in the BC context. Consistent with research on teachers' LAL, we found that teachers' use of the ELL Standards facilitated teachers' knowledge about language teaching, learning, and assessment practice and issues and tensions relating to language and curriculum assessment (Coombe et al 2020).

A key challenge in engaging in these collaborative approaches to professional learning for assessment literacy is time and resources, as these are indicated as key components for starting and sustaining a PLC (DuFour and Eaker 1998). Educators have limited opportunities to engage in targeted professional development outside of the classroom, and educational leaders and administrators have a limited pool of resources to support release time for these activities. However, these collaborative practices with educators across districts, schools, and grade levels, as well as with university researchers, afforded them with many opportunities to understand how to create, apply and provide feedback on standards, and to have the teacher voice heard. The reciprocal, dialogic exchange of ideas and strategies amongst colleagues from various contexts helped them realize that they were not alone in their endeavor to support ELLs, and on the road to developing a sustainable community of practice in this regard.

Overall, we found that teachers were invested in and keen to develop assessment literacy to improve and build their assessment practice. There was

some evidence for potential growth in assessment literacy with focused, strategic professional conversations and learning. The field testing and interactive professional learning supported the valuing of teachers' voices and agency in language education and was an important counter-balance to professional learning approaches where teachers may be passive audience members listening to an expert.

In addition to sharing recommendations from the teachers' field-testing of the standards, critical perspectives emerged from listening to other teachers. Although they were not expected to engage in discussions about the use and implications of standards, and the values and power relations associated with their use, there may be room for further teacher professional learning in this area moving forward. This is supported by the research on critical perspectives in the teacher assessment literacy/competency literature, which highlights the importance of teachers' recognition of students' broad linguistic repertoires and the kinds of ideologies they are bringing to their practice (García and Lin 2017).

In analyzing the teachers' reflections, questions, recommendations, and observations, and by examining the recent research and trends in other contexts, we suggest that in the BC context, next steps require ongoing refinement and improvement of the ELL Standards, based on use in the classroom and impact on students, teachers, and school communities. In particular, the standards should be more inclusive of the sociocultural and linguistic practices of students in BC education, and should explicitly address multilingualism in the classroom. Teachers' language assessment literacy can be addressed and fostered to support inclusive, equitable practice and engage students' language resources in curriculum learning to maximize educational opportunities. These aims represent an ongoing commitment by provincial and district leaders to develop teachers' assessment literacy in an ever-changing linguistic and cultural educational landscape.

Acknowledgements

We would like to thank the following district leaders for their collaboration and support of this project: Ann Hunter, Vancouver School Board; Ann Thorpe, Burnaby School Board; Hieu Pham-Fraser, Richmond School Board; and Maria Yioldassis, West Vancouver School Board.

References

BC Ministry of Education (2017) *English Language Learning (ELL) Standards*, available online: www2.gov.bc.ca/assets/gov/education/kindergarten-to-grade-12/teach/pdfs/ell/ell-standards-full.pdf

BC Ministry of Education (2019) *Education by the Numbers*, available online: news.gov.bc.ca/releases/2019EDUC0069-001642

Coombe, C, Troudi, S and Al-Hamly, M (2012) Foreign and second language teacher assessment literacy: Issues, challenges and recommendations, in Coombe, C, Davidson, P, O'Sullivan, B and Stoynoff, S (Eds) *The Cambridge Guide to Second Language Assessment*, Cambridge: Cambridge University Press, 20–29.

Coombe, C, Vafadar, H and Mohebbi, H (2020) Language assessment literacy: What do we need to learn, unlearn, and relearn? *Language Testing in Asia* 10 (3), 1–16.

Cumming, A (2009) What needs to be developed to facilitate classroom-based assessment?, *TESOL Quarterly* 43 (3), 515–519.

Cummins, J, Jang, E E, Stille, S, Wagner, M, Byrd Clark, J and Trahey, M (2009) *Steps to English Proficiency (STEP): Validation study*, Final research report presented to the Ministry of Education, Toronto: Modern Language Centre, OISE.

Davison, C and Leung, C (2009) Current issues in English language teacher-based assessment, *TESOL Quarterly* 43 (3), 393–415.

DuFour, R and Eaker, R (1998) *Professional Learning Communities at Work: Best Practices for Enhancing Student Achievement*, Bloomington: National Educational Service.

Edelenbos, P and Kubanek-German, A (2004) Teacher assessment: The concept of "diagnostic competence", *Language Testing* 21, 259–283.

García, O and Lin, A M Y (2017) Translanguaging in bilingual education, in García, O, Lin, A M Y and May, S (Eds) *Bilingual and Multilingual Education*, Encyclopedia of Language and Education, New York: Springer International Publishing, 117–130.

Garnett, B W (2008) *A critical examination of academic trajectories of ESL youth*, unpublished PhD thesis, University of British Columbia.

Gunderson, L (2021) The dangers of labels: English as a Second Language. *Research Outreach*, available online: researchoutreach.org/articles/dangers-labels-english-second-language/

Inbar-Lourie, O (2008) Constructing a language assessment knowledge base: A focus on language assessment courses, *Language Testing* 25 (3), 385–402.

Jang, E E, Cummins, J, Wagner, M, Stille, S, Dunlop, M and Starkey, J (2015) Assessing school-aged English language learners in Ontario curriculum learning contexts using Steps to English Proficiency, *Language Assessment Quarterly* 12 (1), 87–109.

Jones, N and Saville, N (2009) European language policy: Assessment, learning and the CEFR, *Annual Review of Applied Linguistics* 29, 51–63.

Li, G, Anderson, J, Hare, J and McTavish, M (2021) *Superdiversity and Teacher Education: Supporting Teachers in Working with Culturally, Linguistically, and Racially Diverse Students, Families, and Communities*, New York: Routledge.

Rust, C, O'Donovan, B and Price, M (2010) A social constructivist assessment process model: How the research literature shows us this could be best practice, *Assessment and Evaluation in Higher Education* 30 (3), 231–240.

Saldaña, J M (2015) *The Coding Manual for Qualitative Researchers* (Third edition), Thousand Oaks: Sage Publications.

Scarino, A (2013) Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning, *Language Testing* 30 (3), 309–327.

Stille, S, Jang, E E and Wagner, M (2015) Building teachers' assessment capacity for supporting English language learners through the implementation of the STEP language assessment in Ontario K-12 schools, *TESL Canada Journal* 32 (9), 1–23.

Talmy, S (2015) A language socialization perspective on identity work of ESL youth in a superdiverse high school classroom, in Markee, N (Ed) *Handbook of Classroom Discourse and Interaction*, New York: Wiley Blackwell, 353–368.

Tremblay, C and Hall, B L (2014) Learning from community-university research partnerships: A Canadian study on community impact and conditions for success, *International Journal of Action Research* 10 (3), 376–404.

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

Volante, L and Fazio, X (2007) Exploring teacher candidates' assessment literacy: Implications for teacher education reform and professional development, *Canadian Journal of Education* 30 (3), 749–770.

3 Self-assessment skills as part of English and French student teachers' assessment literacy

Olga Ukrayinska
Kharkiv Skovoroda National Pedagogical University, Ukraine/Lancaster University, UK

This chapter describes a case study project designed to develop self-assessment skills among English and French student teachers at Foreign Languages Departments in Ukrainian universities. A course on language testing and assessment was introduced for Master's degree students in Kharkiv Skovoroda National Pedagogical University in 2009. This case study describes the updating and trialing of new content of this course related to self-assessment, in three Ukrainian universities across the academic year 2018–19. During the COVID-19 pandemic and the war in Ukraine the course was taught online, which required some adjustment of course provision and put some limitations on methodological interaction between students and their teachers.

The chapter focuses on:

- A discussion of the benefits of self-assessment competence, as an element of language assessment literacy (LAL)
- A detailed description of the trialing of the new course content, which I termed 'experimental teaching', to foster such competence as part of the LAL development of Ukrainian foreign language pre-service teachers. This description includes the context of the training, self-assessment methods, strategies, tools and activities that enabled them to assess the outcomes of their own learning
- A discussion of some of the results of a case study examining the possible benefits of the experimental teaching

У главі описано досвід формування компетентності з самооцінювання у майбутніх вчителів англійської та французької мов в українських закладах вищої освіти. Курс з мовного тестування та контролю для студентів-магістрантів було запроваджено у Харківському національному педагогічному університеті імені Г.С. Сковороди у 2009

році. Оновлений зміст курсу апробовано під час експериментального навчання у трьох українських закладах вищої освіти у 2018–2019 навчальному році. Під час пандемії COVID-19 та війни в Україні курс викладався онлайн, що передбачало зміни у викладанні курсу та наклало певні обмеження на методичну інтеракцію між студентами та їх викладачами.
Глава фокусується на:

- Перевагах розвитку компетентності з самооцінювання як елементу методичної контрольно-оцінювальної компетентності майбутніх вчителів англійської та французької мов.
- Процесі формування компетентності з самооцінювання під час експериментального навчання майбутніх вчителів англійської та французької мов контролю та оцінюванню. Описано організацію навчання, методи, стратегії і засоби навчання, надано приклад завдань з самооцінювання, за допомогою яких студенти оцінювали результати своїх навчальних досягнень з оволодіння англійською та французької мовами і контрольно-оцінювальною діяльністю.
- Результатах експериментального навчання і перспективі використання запропонованої методики формування компетентності з самооцінювання для розвитку методичної контрольно-оцінювальної компетентності майбутніх учителів іноземних мов

Introduction

For almost two decades self-assessment has been considered an indispensable part of Foreign Language (FL) teachers' assessment literacy since the Common European Framework of Reference for Languages (CEFR) was launched in 2001 by the Council of Europe. This competence comprises knowledge, skills and the ability of student teachers to teach learners self-assessment as well as to assess their own language proficiency and the quality of their professional activities, including the efficiency of their assessments. Huang (2022) worked with English teachers whom he engaged in self-assessment and discovered that this experience helped the participants not only boost their own self-efficacy and self-confidence but also that of their learners. This was achieved by reducing task difficulty, improving learners' skills and helping them to become self-regulated. Engaging students in self-assessment includes motivating them to practice it, explaining how to do it and guiding them by providing feedback and appropriately exploiting its results. For this, teachers need to become experts in self-assessment themselves, which means they need to be taught about self-assessment

pre-service and trained to practice it through a set of subject-specific activities under their teacher trainer's supervision.

Self-assessment is closely related to and associated with self-appraisal, self-reflection, self-monitoring, self-regulation and self-evaluation (Brown and Harris 2014, Martinez, Mon, Alvarez, Fueyo and Dobarro 2020, McMillan and Hearn 2008, Wride 2017). Despite the differences and nuances of these terms, all of them imply accountability of students for their learning. The development of self-assessment skills is a challenging task which requires continuous efforts by teacher educators and prompts researchers' attempts to find appropriate tools. Brown and Harris (2014:23) argue that 'self-assessment is teachable and learnable' if done systematically as a curricular competence. Teachers should spend time preparing their students for self-assessment (Brown and Harris 2014, Wride 2017), notably deciding upon success criteria, clarifying standards and learning outcomes (Wride 2017). Students should also be taught about pertinent self-assessment strategies and activities. Consequently, the roles of a teacher are facilitator and moderator of assessment in collaboration with students (Martinez et al 2020, Wride 2017).

While there is much research dedicated to self-assessment by FL learners, there is little dedicated to teachers' self-assessment of their instructional activities and none to developing self-assessment skills of student teachers. The question arises here of how to organize the corresponding pre-service training and which materials to use.

Theoretical background

Self-assessment has been implemented in tertiary FL education to promote students' learning autonomy. It builds upon the student-centered learning approach, included in curricula to better address particular learners' needs and interests and stimulate peer- and self-learning. Self-assessment allows students to adjust a lesson to their pace, and to support their problem-solving and ability to use feedback to continue their development (Keiler 2018). The idea of practicing self-assessment is to actively engage learners in their FL acquisition, to develop their capacity for self-direction, and to redelegate responsibility for the results of their studies to them (Harrison, O'Hara and McNamara 2015, McMillan and Hearn 2008, Martinez et al 2020).

McMillan and Hearn (2008:40) define self-assessment as 'a process by which students monitor and evaluate the quality of their thinking and behaviour when learning and identify appropriate strategies that can improve their understanding and skills' through identifying discrepancies between their current and desired performance. Wride (2017) and Tai, Boud, Dawson and Panadero (2018) note that gradual development of evaluative judgment enables students to make decisions about achieving learning outcomes. Self-assessment increases students' motivation and interest in

studying, which leads to enhanced learning and development of their critical skills (McMillan and Hearn 2008, Wride 2017). However, in order to be successful it should be purposeful and systematic (Wride 2017). Bailey (1981) defined teacher self-assessment as the process of self-examination for the purpose of instructional self-improvement which will enable the teacher to become aware of personal classroom teaching effectiveness, to learn how to control classroom instructional behaviours and to become self-directed in instructional improvement activities. For the purposes of this project, self-assessment refers to formative assessment by trainee language teachers of their own FL performance as well as their assessment practice.

There are various interpretations of the terms related to self-assessment, namely 'methods', 'strategies', 'tools', and 'activities'. For instance, reflection journals can be called 'a method', 'a strategy', 'an activity'. In order to avoid confusion I made an attempt to distinguish these terms to fit my learning context.

Firstly, the following self-assessment *methods* are distinguished: observation, discussion, questioning, project work, product development, analysis of audio records and videotapes, and reflection on the recent performance.

Secondly, a self-assessment *strategy* is a practical implementation of a self-assessment method or technique depending on the student's personality, their goals/objectives of learning a FL, their strengths and weaknesses, the level of their familiarisation with appropriate techniques or some external factors. For example, a student may choose to video record their response, take notes while responding or just rely on their memory; they may decide on how to address a problem – to read grammar reference books, to learn the rules by heart, to consult the teacher or to do a series of corresponding exercises. A self-assessment strategy is an appropriate way to collect information, to evaluate the product depending on the task and the response, to order steps of achieving the learning goal, to monitor progress towards it, to keep track of individual progress and to receive feedback. Students need to develop a habit recognizing their unique needs, otherwise random actions will not lead to success. Self-assessment strategies include reflective writing, listing strengths and weaknesses or systematic mistakes, observing one's performance or building up a portfolio.

Thirdly, self-assessment *tools* are instruments used to measure individual progress towards academic proficiency in a specific subject area: (grading) rubrics, self-assessment open-ended or short prompt questions, questionnaires, scripts, checklists, reflective journals or logs, portfolios, progress cards, emoji worksheets, structured interviews, rating scales, multiple-choice tests, true–false statements and surveys. There are numerous expert-developed tools such as self-assessment European Language Portfolio (ELP) grids (Little and Perclova 2001), self-assessment

descriptors in *European Portfolio for Student Teachers of Languages* (Newby et al (Eds) 2007) or descriptive scales in the CEFR (Council of Europe 2001) or its later versions. Apart from these examples there are numerous achievement charts and grids taken from a Google search offered to English and French learners for self-assessment purposes.

Finally, self-assessment *activities* provide student teachers with an opportunity to make judgments about their learning process and products, based on success criteria that they have agreed upon with their teacher. Self-assessment activities are context-specific and task-based. Bailey (1981) recommends focusing on a small number of skills at a time, practicing self-analysis systematically.

Experimental teaching

The student-teacher participants

The experimental teaching was conducted with 108 Ukrainian students from Kharkiv Skovoroda National Pedagogical University, Kharkiv Karazin National University and Kharkiv University of Humanities 'People's Ukrainian Academy'. There were 99 female students and nine male students in the cohort, with an average age of 21–22. The participants were at the final stage of their studies and their majors were Teaching English or French as a Foreign Language. Whilst most of them have part-time jobs, less than 10% work in the education sector. According to the survey done prior to the course they appeared not to have much experience in self-assessment except for checking their answers against keys during their university classes.

Details of the experimental teaching

The course content was based on the curriculum for Foreign Language Development, European reference documents such as the CEFR (2001) and *European Portfolio for Student Teachers of Languages* (Newby et al (Eds) 2007) and multiple Internet resources (achievement rubrics and checklists). Teaching was carried out in English for English student teachers and in French for French student teachers correspondingly.

The course consists of a number of classes of lectures, instructor-fronted classes focused on presentation of more theoretical content. These classes are followed by more seminar-style meetings with more student-centered interactive tasks. The course also has a number of asynchronous activities completed outside of class meeting times. During the first lecture of the course the students were provided with guidelines on completing the reflective journal, so that they could collect two months' observations of their own learning process. A subsequent lecture included self-assessment

methods, strategies, tools and activities. In this lecture, students were also provided with justification and guidelines for compiling their portfolios.

Two seminars of the course were dedicated to practicing self-assessment. The standards of teaching FL at Foreign Languages Departments in Ukraine were introduced to the students. Based on these documents, assessment criteria were elaborated jointly with the view of using them in future teaching practice. After this cooperative work, the students proceeded to pair work practicing peer-assessment, and also watched some videos showing people doing self-assessment. Subsequently, the students tried their hand at self-assessment in the classroom using ELP grids. Finally, the students recorded their oral performance and drafted a series of written works meant to be self-assessed against the checklists provided by the lecturer. Throughout the course the students were asked to keep their reflective journal. They were expected to write the journal in free style with the lecturer's guidelines in mind; however, they demonstrated lack of understanding of what exactly to write. Due to this, I prepared 19 prompt 'Yes/No' and 'Wh-questions' for them to scaffold their work. Journal entries also included comments on how their language abilities were being assessed in their General English or French courses. Individualized learning goals were set depending on difficulties experienced by the students based on the feedback given by their teachers and reflection performed by the students in their journals.

Below is the summary of the activities and tools used in the experimental teaching to develop the students' self-assessment skills in terms of a) their assessment literacy and b) their own FL learning:

- conducting literature analyses, reading about curriculum requirements and analysing the corresponding assessment criteria (length of texts for reading; duration of audiotexts; task types; communicative functions)
- developing rubrics with a certain task in mind
- discussing assessment criteria of the sample videos provided by the Council of Europe corresponding to the CEFR levels and the recorded speech or writing samples of Ukrainian students
- simulating assessment activities such as item-writing, marking and rating
- reviewing and editing by the students of their own writing, their own items and rating scales with the use of checklists provided by the lecturer
- providing open-ended comments on course documents as well as in oral interviews and surveys.

Apart from these activities, as their lecturer I observed the students doing self-assessment, asking them questions and taking notes, and I evaluated their self-assessment materials to document improvements during the course.

During the teaching of the course, I observed that in addition to needing more guidance to complete their reflective journals, these MA students also needed more scaffolding to perform a deep reflection on their results in the course against the curriculum expectations. Therefore, I provided them with simplified checklists corresponding to their language development as well as their skills in conducting assessment. Examples include: providing a self-assessment checklist for their formal letter writing assignment and also for their drafts of their oral interaction rating scale (for more details see Ukrayinska 2020).

Here is one example of an assignment to practice self-assessment of the students' language abilities: 'Write an argumentative essay. Self-mark it using the checklist. Define typical drawbacks, develop a plan to eliminate them. Edit your essay.' After that the students exchanged their essays with their groupmates and practiced peer assessment. The essays written on the same topic were edited again in line with the feedback given by peers. All the versions of essays and feedback were then submitted to the lecturer.

Results

Here are insights from the experimental programme and a few selected results and thoughts about FL student teachers' self-assessment skills in LAL development. Firstly, in accordance with the curriculum for Foreign Language Development, students in their fifth year are expected to be at C2 level of the CEFR and lecturers attest that most of them appear to be functioning at this level. However, the results of self-assessment done against the 'Can Do' statements (Little and Perclova 2001) at the introductory seminar showed that 82% of them underestimated themselves, choosing Level B2 or even B1. They justified their choices by saying that the descriptors of C1 and C2 levels mention some things they do not do in their second language (e.g., writing letters, reading contracts, listening to public announcements, particularly at stadiums). For that reason, the students did not appear to be confident enough to assign themselves a higher level. Hence, the students were recommended to focus not on the scope of activities they do but on their quality.

When surveyed at the beginning of the course, 91% claimed that it is not possible to self-assess objectively and that it is their teachers' job. This view was also reflected in their responses in the reflective journals, where 86% of them strongly agreed with marks their teachers gave them. This might mean that when the students self-assessed their work before, their teachers collected the results and gave them marks but did not teach them to reflect and learn from mistakes. As a result, it was challenging to convince the students that they can benefit from self-assessment as a component to the results of the teacher assessment.

When I designed the content for self-assessment module, one of the challenges was to find appropriate tools. For this I analysed a number of achievement charts and grids. Achievement charts can contain smileys or emojis, approving and disapproving hand gestures (e.g. thumbs up) or the options 'Yes/No', 'Oui/Non', which I considered to be more appropriate for young learners. Achievement grids can contain descriptors of a particular skill, for example, 'Can Do', 'Je peux' statements. Rubrics of some checklists under analysis were too generalized and not product-oriented (e.g. 'I tried really hard'; 'I spoke clearly'; 'I can use transition words to organize my ideas in writing', 'The story makes sense to me'; 'J'ai justifié mes opinions', 'J'ai employé le vocabulaire de la description'), consequently it can take students a lot of time to internalize their content. My assumption is that it is more effective to start with task-specific rubrics and then proceed with generalized samples. For instance, it can be a checklist for the argumentative essay 'Do foreign language teachers' questions violate learners' privacy?' containing criteria (content, coherence/cohesion, range of grammar, range of vocabulary, structure, spelling, length) with clear detailed rubrics. As for reflection such rubrics as 'What I did well was …' or 'Trois points que je dois améliorer', in my subjective opinion, are not likely to be informative unless self-assessment has been extensively practiced and students have a clear understanding of the respective outcome (*what should be done?; how well should it be done?; how can I do it as well as possible?*). Bearing these assumptions in mind I developed a number of checklists described above and engaged students in developing task-specific rating scales.

I distinguish several groups of activities depending on the stage of teaching self-assessment: a) activities for familiarization with standards: 'retrieve the components from the rating scale/descriptor', 'define the key words in assessment criteria', 'classify the descriptors into criteria', 'answer the questions', 'watch/listen to sample performance and assign points to it', 'rank the works'; b) self-assessment activities: curriculum mapping, summarizing, hand signaling, grading yourselves, self-marking, completing checklists, doing multiple-choice or true–false tests, answering prompt questions, writing reflective journals, checking answers against the key, editing pieces of writing or scripts.

It is necessary to mention that it is often not possible to see immediate results of practicing self-assessment as it takes time and has long-term goals. However, such activities as self-checking a piece of writing, applying a verbal protocol when doing a reading task and using checklists proved to motivate the students; they reported using some of these tools in their courses long after they were required to do so. For example, the students reported that they found it helpful to underline, circle, or tick elements in their work under consideration or mark them with pencils of different colours or electronic highlighting.

In addition, they demonstrated improvement in their coursework. For example, by using checklists for self-reviewing the items and rating scales the students learned to avoid missing structural components of items, to avoid grammatical cues, and to keep options and distractors homogenous. Below is a sample of modifying an item against the checklist.

Original item	Modified item
3) Gapyear.com was founded **A.** as a platform for finding the cheapest holidays packages **B.** as a social network where backpackers could share their stories and experiences **C.** as a blog where Tom Griffiths shared his experiences and gave advice to the masses on backpacking and travelling	**3.** *Gapyear.com* was founded … **A** as a travel booking platform. **B** as a social network for backpackers. **C** as a blog of travelling students.

Even though we collected rich information during this case study, the analysis of the students' reflective responses showed that there is a need to change some questions or add follow-up questions, for example, the question 'Which task formats do I feel to be most difficult/easiest?' needs to be supplemented by the question 'Why?' to learn more about the strategies of doing such tasks since some students did not give exhaustive answers. Also the question 'Do I know how I progress?' should be accompanied with a request for evidence. Mere answers of 'Yes' or 'No' are not informative and cannot contribute to understanding their assessment literacy development.

Conclusions

The empirical data collected during the experimental teaching demonstrated that developing self-assessment skills can be carried out in class as well as online, which is especially valuable under the war circumstances in Ukraine when not all students have an opportunity to join synchronous online classes. Students need to be provided with detailed guidelines to work asynchronously when developing their LAL.

The experimental teaching helped shape instructional materials and fix procedures of developing self-assessment skills as part of LAL development, even though the time allocated to these activities in the course was limited. In addition it contributed to development of my own assessment literacy, posing challenges and showing ways of responding to them. I learned the necessity to separate the order of taught aspects: assessment – peer

assessment – self-assessment. I also learned that it would be preferable to introduce a needs analysis before doing self-assessment with the 'Can Do' statements. From what I learned from this experience, I modified the questions in the journal, and specified some points in the checklists for evaluation of items and scales.

Self-assessment provided motivation to the students to enhance their language proficiency and assessment literacy. It also activated them to engage with assessment, their future professional activity. However, this case study revealed some resistance to self-assessment, and some doubts about its usefulness. This is an important challenge to address in future work: if students can be convinced that self-assessment can be a powerful instructional tool, they will be more likely to encourage their learners to undertake self-assessment.

There is no doubt that teaching self-assessment is a time-consuming but promising process. It should be done throughout university studies in order to build a self-assessment habit and to provide future teachers sufficient time to discover and address weaknesses and to observe and reflect on assessments carried out by them and on them.

References

Bailey, G (1981) *Teacher Self-Assessment: A Means for Improving Classroom Instruction*, Analysis and Action Series, Washington, D C: National Education Association.

Brown, G and Harris, L (2014) The future of self-assessment in classroom practice: reframing self-assessment as a core competency, *Frontline Learning Research* 3, 22–30.

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

Harrison, K, O'Hara, J and McNamara, G (2015) Re-thinking Assessment: Self- and peer-assessment as drivers of self-direction in learning, *Eurasian Journal of Educational Research* 60, 75–88.

Huang, Q (2022) Influence of EFL teachers' self-assessment on their self-regulation and self-efficacy, *Frontiers in Psychology* 13, 1–8.

Keiler, L S (2018) Teachers' roles and identities in student-centered classrooms, *International Journal of STEM Education* 5 (34), 1–20.

Little, D and Perclova, R (2001) *The European Language Portfolio: A guide for teachers and teacher trainers*, Strasbourg: Council of Europe.

Martinez, V, Mon, M, Alvarez, M, Fueyo, E and Dobarro, A (2020) e-Self-Assessment as a strategy to improve the learning process at university, *Education Research International* 2020, 1–9.

McMillan, J H and Hearn, J (2008) Student self-assessment: The key to stronger student motivation and higher achievement, *Educational Horizons* 87 (1), 40–49.

Newby, D, Allan, R, Fenner, A-B, Jones, B, Komorowska, H and Soghikyan, K (Eds) (2007) *European Portfolio for Student Teachers of Languages: A Reflection Tool for Language Teacher Education,* Strasbourg: Council of Europe.

Tai, J, Boud, D, Dawson, P and Panadero, E (2018) Developing evaluative judgment: enabling students to make decisions about the quality of work, *Higher Education* 76, 467–481.
Ukrayinska, O (2020) Using checklists for developing student teachers' language assessment literacy, in Hidri, S (Ed) *Perspectives on Language Assessment Literacy*, New York: Routledge, 84–106.
Wride, M (2017) *Guide to Self-Assessment*, Dublin: Academic Practice, University of Dublin Trinity College.

4 EFL students' knowledge and perceptions of international English tests: A text-mining approach in social media

Mojtaba Heydari
University of Ottawa, Canada

Fahimeh Marefat
Allameh Tabataba'i University, Iran

This chapter provides an example of the application of text mining to test-taker discussions of large-scale testing, specifically of International English Language Testing System (IELTS) candidates in an IELTS preparation social media group, in order to:

- Understand the key topics discussed in the group
- Examine the learners' knowledge and perception of the IELTS test
- Identify the aspects and sections of the test which learners appeared to regard as most problematic

در این فصل با بهره‌گیری از روش متن کاوی و تحلیل داده‌های انبوه، پیامهای ارسال شده‌ی داوطلبان آزمون آیلتس در یکی از گروههای تلگرامی محبوب آمادگی آیلتس مورد بررسی قرار گرفته شده است. با بررسی بیش از بیست و شش هزار پیام کاربران در قالب سوال و گفتگو در این کانال، موارد زیر مورد مطالعه قرار گرفتند:

- موضوعات کلیدی مورد بحث در گروه
- دانش و درک فراگیران از آزمون آیلتس
- مشکل سازترین جنبه های آزمون

یافته های این مطالعه می تواند به بهبود درک ما و ذینفعان آزمون (سیاستگذاران، دانشگاهها، دولتها، برگزار کنندگان و سازندگان آزمون) از تلقی زبان‌آموزان از آزمون آیلتس و سایر آزمونهای بین‌المللی مشابه کمک کند. با توجه به کمبود مطالعات با تمرکز بر سطح دانش زبان‌آموزان در مورد آزمون‌ها، این مطالعه می تواند جهت نگاه ما را متمرکز به دغدغه‌های زبان‌آموزانی کند که برای موفقیت در این آزمون تلاش می‌کنند.

The findings from this study can potentially help improve both stakeholders' and our own understanding of the test-takers' perceptions of IELTS and other similar international tests. Given the scarcity of studies focusing on the learners' knowledge of tests, this study can provide a voice to learners who strive to succeed in such tests and help us understand their impact on their lives.

Introduction

More and more people are exchanging information on social media platforms such as Facebook, Twitter, Instagram and Telegram. Due to the vast volume of data posted, these platforms contain much untapped potential for extracting knowledge and analysing the understandings of their users. Text mining, an automated technique that uses computational algorithms to extract meaning and patterns from text (He 2013), identifies units of meaning and relationships that would otherwise remain buried in the mass of textual big data. Despite its versatile potential, the use of text mining in education and language learning has only recently emerged (Warschauer, Yim, Lee and Zheng 2019). While to date many studies have explored issues related to high-stakes tests like IELTS from theoretical as well as empirical viewpoints, test-takers' perceptions of such tests and language assessment literacy (LAL), concerning broader conceptions of fairness and justice, have not been given adequate attention in the literature (Hamid, Hardy and Reyes 2019). In addition, social media, an invaluable resource for delving into the test-takers' perceptions, have been largely neglected; instead, test-taker attitudes have been mainly investigated by means of interviews and questionnaires. Moreover, in conjunction with social networks, educational data mining can exploit genuine data produced in learners' everyday lives to gather insights that cannot be gained otherwise, and to enhance teaching, learning and decision-making (He 2013).

IELTS in the context of Iran

IELTS is a high-stakes English test that evaluates a candidate's ability in four skills: listening, reading, writing, and speaking. Academic and General Training versions of the test are available, with the former being used for tertiary admission and the latter for immigration purposes (Pearson 2019). In the past decade, Iran's young population has shown an unprecedented tendency to take high-stakes tests like IELTS to qualify for emigration to English-speaking countries or for admission to foreign universities. This tendency is triggered mainly by their country's economic and political problems, which drive them to seek free, economically stable, and democratic destinations for emigration. IELTS is the most widely used test

in Iran, as the IELTS score is increasingly accepted by academic institutions and governments worldwide for educational or immigration purposes (Saif 2021). Iranian students are among the top 25 nationalities worldwide taking the test annually, describing themselves as 'highly motivated' and perceiving 'the test score as consequential to their futures' (Saif 2021:509). In the recent decade, the increased exchange rates and the economic turmoil in Iran have made the test very expensive, almost equal to the monthly minimum wage in Iran (Financial Tribune.com), not counting preparation expenses. This cost has effectively heightened the stakes of the test for students who feel more pressure to do well on the test at first sitting and who may not be able to afford another chance.

In this study, we employ text mining to provide first-hand insights into test-takers in a large English as a Foreign Language (EFL) context in order to understand the key topics discussed by the learners and examine their knowledge and perception of the IELTS test. We also aim to identify the sections of the test (listening, reading, writing, speaking) that learners perceived to be most challenging, as previous research has shown jagged score profiles, potentially indicating variable levels of comfort across test sections (e.g. Pearson 2019). The findings of this study can give voice to the test-takers and, ultimately, provide us and the stakeholders with a deeper understanding of the test-takers' perceptions about such international tests from socioeconomic, political, and ethical perspectives.

Methods

Data and participants

In Iran, numerous interest groups and channels exist on social media to help the (potential) test-taker share and find relevant information about IELTS and help them prepare for it. We targeted one of Iran's most popular Telegram groups, founded by a well-known IELTS preparation institute, with 22,889 members on the date of data collection. The group's administrator is an English teacher, and the other members are primarily students aiming to take part in the IELTS exam or improve their existing marks. While the learners are the core content creators, some occasional ads are posted by the institute owning the group. To collect the data, we covered a period of six months, from 1 September 2021 to 28 February 2022. This yielded a total of 21,622 messages, totalling 371,898 words. In the process of data collection, we filtered out all the sponsored content and ran the analysis for the posts made by the learners. Among the 22,889 members of the group, 8,211 members (36%) were active during the six months of data collection by posting at least one message. It should also be noted that collecting the biodata of the users is not possible due to their anonymity.

Data processing and analysis

A Python 3 ready-made web-crawling application was used to analyse the data. Since most texts on platforms like Telegram are usually loosely organised and unstructured, it is important to pre-process the text (i.e., clean the data) before it can be analysed. This step was challenging because although the default language of the channel and most posts was Persian, some posts were (completely or partially) written in English. In addition, there were posts written in Pinglish (Persian words written using Latin alphabet or English words written using Persian alphabet) which needed to be re-written for running the analysis. This was done by automatically deleting meaningless words (i.e. words that were so poorly written or spelt that they could not be identified), correcting misspelt words, and manually rewriting Pinglish words into English. Also, the text was normalised by: 1) unifying the characters and replacing non-standard characters with standard ones, e.g. آ instead of ا in a word like آزمون; and 2) normalising spaces and non-breaking spaces, e.g. non-breaking spaces (نیم فاصله) in Persian can make problems in counting the frequency of words like دانش‌آموز.

In order to identify the key topics, an automatic topic modelling using a classification algorithm with words randomly assigned under different topics was employed (Namugera, Wesonga and Jehopio 2019). The key topics were then manually reviewed to identify the overlaps or the words that were incorrectly assigned to the topics.

We ran a keyword analysis to identify the keywords of the text related to the four skills. It should be noted that since many keywords were written in both Persian and English, we ran separate word counts for both languages and the numbers written in the tables show the sum of the keywords written in both languages.

Results

Key topics

We came up with the following key topics:

Grammar and translation. Several questions or comments were posted to ask for grammar tips or whether a specific sentence was grammatical or not. An interesting issue was that some students used technical grammatical terms in their questions, which was a sign of preparing for IELTS: *'Hey guys! A question! Can I reduce a non-definitive relative clause or not?'*[1]

1 All the quotes reported in this study were originally written in Persian and were translated into English by the authors. To access the original quotes, you may contact the authors.

Also, a few learners asked for translation tips from Persian to English and vice versa. This mainly happened for phrases common in the IELTS test yet not very frequent in the students' everyday use (due to the differences in their background knowledge and culture): *'What's the translation of political correctness?'*

Study resources. A sizeable portion of the posts was related to asking for resources and study materials, including successful writing and speaking sample models. In addition, many learners were also wondering whether certain books were a good match for their proficiency level, target score, time available, etc. Some of the posts were also related to digital resources, including websites, YouTube channels, computer software, and mobile applications:

- *'Is there a difference between the third and fourth editions of Essential Words for IELTS?'*
- *'Does Merriam-Webster dictionary have an Android application?'*
- *'I'm looking for good writing samples! Are Rachel Mitchell's samples good?'*

General guidance. Students were looking for general information about preparing for and taking the IELTS test, including preparation time, suitable courses, study methods (e.g., self-study versus attending a class), costs, and asking for good teachers/institutes.

- *'In my opinion, skills-based workshops are more useful than IELTS classes. IELTS classes are good for those in a hurry.'*
- *'From which level did you start, and how long did it take you to get to this mark?'*
- *'They give you a shock in the listening and reading section, but you should keep calm because that's the nature of IELTS. Expect something more difficult than Cambridge samples, so you won't be shocked. For the listening section, practice Actual Listening and increase the speed of Cambridge listening.'*

Surprisingly, there were numerous tips and comments which included technical terminology often used by teachers. Using keyword analysis, we found the following technical words used by the learners to ask for and give recommendations: *skimming, scanning, shadowing, body paragraph, thesis statement, main idea, mind map, fluency, accuracy, idiomatic expression, active/passive vocabulary, band descriptors, complexity, coherence,* and *topic relevance.* There were examples of the sentences which were automatically assigned to the 'study resources' or 'grammar and translation' topics but we re-assigned them to the 'general guidance' topic after we manually reviewed the sentence.

Exam logistics. We found that there is a lot of interest regarding exam equipment, computer-delivered vs paper testing, registration, required documents, timing, location, etc.

- *'I believe in the real test you use headphones.'*
- *'I preferred centre … in terms of comfort and atmosphere. Especially their headphones and chairs; I had no headache or backache after the test.'*
- *'Is there a website or software to practice computer-based test? Please give me a link!'*

Severity/leniency of examiners/centres. One of the hottest issues discussed in the channel was related to the severity and leniency of examiners or centres. There seems to be a general belief among the students that some examiners are unfair or more severe than others, and some exam centres are generally fairer. This has initiated a discussion strand among the students to introduce and identify more lenient examiners and exam centres based on their experience.

- *'In the real test day in … test centre, the speaking examiner was Mr. …. He asked me a lot of follow-up questions but gave me a 7. It was a very good mark compared to two tests and many mock tests. I'm thankful to him.'*
- *'For sure, … test centre gives you the best and closest marks to your mock tests.'*

Looking for speaking partners. The students had a system to show their interest to practise English with a partner. They used a form to post an ad on the channel to look for a partner by giving information related to their level, available hours, gender (and the preferred gender of their partners), etc. Some of them also specified that they were interested in practising a specific book or online resource.

Complaints/request for remark. The test-takers who are not satisfied with their results may apply for a remark. This usually happens for speaking or writing skills, where the scores are considered subjective. According to test regulations, a candidate may apply for remarking by paying an administrative fee of approximately GBP 60, which is reimbursed should a score be upgraded (Pearson 2019). Due to the high cost of this option compared to retaking the test, candidates who are not satisfied with their results might make a rather tricky decision whether to retake the test or ask for a remark. For this reason, many students use this channel to learn from the experience of other candidates to make this decision. *'I got 8.5 in listening, 9 in reading, 7.5 in speaking, but 6 in writing. Do you think I have any chance to ask for a remark?'*

In response to this question, some students recommended that the candidate applies for a remark since many candidates have previously shared

their successful remark experience. However, there was disagreement among the students whether asking for a remark would change in their favour or not.

> *'Many experiences about asking for remarks shared in the group are positive 😅 It might cause a misunderstanding. I believe there are many people with "unchanged remark" results. It's just they are not as happy as the others to share it.'*

Expressing emotion/looking for support. Some students expressed their feelings of sadness, grief, failure, incapability, and lack of self-confidence as a result of failure in the IELTS test by posting on the channel. It seems that the students shared their voices to look for social support by hearing positive words, finding people with the same problem, or relieving their stress.

- *'Difficult days of waiting 😩'*
- *'He gave me a 6 in the mock speaking. I talked very well; that was unfair. I'm now afraid even to talk.'*
- *'So, let's keep hopeful. You did a big favour. Your positive energy is awesome. Good luck with your test on Sat.'*

Comparing actual test vs mock test results. The students had a strong desire to know whether their actual test results would be similar to their mock test results.

> *'I had practised all the Actuals and Cambridge samples, but I never got under 6.5. I almost got 7 in most cases. However, I got a 6 in the reading; it was much harder. Cambridge listening tests are generally easier; I always got a 7 or 7.5, but I got a 6 both in the mock and real test.'*

Confusion over levels. This topic included confusion over the levels of the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001), benchmarking IELTS scores to other high-stakes tests, preparation courses, and study materials. In Iran, while private institutions widely use the CEFR levels, the classic placement terminology (e.g., pre-intermediate, intermediate, etc.) is still dominant in the everyday use of the learners. This is reflected in our findings presented in Table 1, which shows the total number of words related to the English proficiency level used in the group.

In addition, we observed that many learners seem to be incapable of interpreting the meanings behind these levels. This can partly stem from inappropriate use of levels by private institutes (i.e., assigning higher-level labels to the classes to encourage more student registration), ambiguously named coursebooks, and general misunderstandings of levels.

- *'I'm not sure if IELTS 7 is equivalent to intermediate or pre-intermediate.'*
- *'What is the equivalent IELTS score for 93 in TOEFL?'*
- *'My level is Upper-intermediate; should I start from Cambridge 7 or 11?'*

For example, in the last excerpt, the student thinks that sample test books are named based on their difficulty level (i.e. Book 1 being the easiest and 15 being the most difficult); however, the books are named in the order of publication (i.e. Book 1 is the oldest and 15 is the most recent one).

Table 1 The total number of English-proficiency-level-related words used in the group

Level-related words	Total number of uses
A1, A2, B1, B2, C1, C2	1,012
Beginner, Basic, Elementary, Pre-Intermediate, Intermediate, Upper-Intermediate, Advanced	3,726

Hashtags

The # (or hash) symbol was used to mark topics and searchable keywords. By adding a hashtag to a social media post, users can emphasise an idea while informing others about what they are discussing (Watson 2020). In addition, they can follow conversations related to a hashtag and choose the topics of their interest. Many Telegram group members use hashtags to talk about issues that are not available in traditional contexts, identify and bond over problems that are common to their peer group, and construct knowledge with like-minded users (Gleason 2018). Overall, 85 unique hashtags were used in the group with an average frequency of 25. Table 2 shows the most frequent hashtags (i.e. used more frequently than the average).

As shown in Table 2, while the topics of language skills like writing and speaking are among the most frequently used hashtags, some other hashtags such as #mock, #resell, #Simon, #remark(request) and #Englishpartner are not usually reported in most formal venues like textbooks and journal papers due to the informal nature of such topics. We found that the students rely on mock tests (unofficial exams that simulate the actual test at a much lower cost) as a means of checking their test preparedness. They also rely on websites like IELTS Simon or IELTS Liz, which they find more helpful in improving their skills than many official textbooks. Many students who are active in the Telegram group believe that because Simon is an ex-IELTS examiner, he can give them insider knowledge. For example, a few students in the group quoted Simon on whether it is good from an examiner's point of view to use so-called 'engineered' words (less frequent compound words which may be more likely to receive higher scores, e.g. 'anti-vaccine parents',

'target-related bonus'). Also, issues related to requests for a remark, reselling their in-person mock tests (when they cannot attend the test themselves), and looking for an English partner were among the most commonly used hashtags.

Table 2 Frequency of the most common hashtags in the group

Hashtag	Frequency
#writing	584
#mock*	320
#reading	318
#speaking	238
#resell*	177
#Simon	128
#remark(request)*	70
#Englishpartner	36

**Note: The marked hashtags are translated from Persian, while the rest were used in English originally.*

Keywords

As mentioned earlier, the aim of keyword analysis was to investigate the frequency of keywords related to the four skills of listening, reading, writing, and speaking to find out which of these skills are more challenging for the test-takers. Table 3 shows the total number of each exam skill used in the messages. As shown in the table, writing was the most discussed skill, followed by speaking, reading, and listening. Based on the discussions related to writing, we found that the learners were primarily worried about their writing scores, remarking their writing, ways to improve their writing, tutors who can give feedback on their writing, etc. While the discussions were generally similar for the speaking skill, the topics discussed related to reading and listening concerned preparation resources (mostly books).

Conclusion

Employing the data-mining approach to one of the most popular IELTS channels in Iran gave us access to first-hand data to study the test-takers as a major group of stakeholders in a high-stakes test like IELTS. It provided us with insights not just into test-taker perceptions, attitudes and behaviour, but also into the contextual characteristics of the situations in which their learning and test-taking experience take place.

In our study, we found that IELTS test-takers were dependent on the emotional support, assistance, and feedback of their peers and experts. For this reason, they used this channel to ask for preparation tips and resources,

Table 3 Frequency of the most common hashtags in the group

Skills	Frequency of use
Writing	21,661
Speaking	17,020
Reading	15,294
Listening	10,011

share their learning/test-taking experience, and air their concerns regarding the fairness and validity of the test. They also took advantage of this venue to address one of the most prevalent issues of learners in EFL contexts: a lack of opportunities to practise speaking (Shih 2007). Based on our findings, the test-takers were very concerned about the IELTS scores for their speaking and writing skills. The students' tendency to ask for advice regarding a remark as well as their discussions of severe versus lenient examiners and test centres are evidence of a lack of trust in the reliability of the assessment of these two skills. This finding is in line with those of Pearson (2019), possibly casting doubt on the strength of 'attitudinal research that has indicated that candidates generally hold favourable perceptions towards trust and fairness in IELTS' (2019:13). One of the factors that may account for test-takers' increased anxiety is the test's particularly high stakes in developing countries like Iran, where economic constraints make retaking the exam difficult to afford. The negative comments about test fairness found here could be explained by their doubts over single examiner marking, confusion over the rating criteria, inconsistency between the mock test and actual test results, and the lack of feedback provided after the test. This resonates with previous findings indicating that many test-takers experience substantial variations in scores across test sittings when they repeat their tests (Hamid et al 2019).

The fact that test-takers are questioning the credibility of IELTS marks suggests that IELTS's regular publication of documentation demonstrating the test's reliability and validity is not sufficient, possibly because such documentation is often released for academics and specialists rather than test-takers. As discussed by Hamid et al (2019:2), 'while scholars and researchers engage with fairness, justice and validity in an intellectual sense, test-takers experience the consequences of different degrees of (un)fairness, (in)justice and (in)validity in very material ways'. Previous research has shown that test-takers experience a lack of fairness and a sense of injustice that can raise concerns about the technical excellence associated with IELTS (Hamid et al 2019).

Based on these findings, IELTS should invest more in finding more efficient ways to improve students' test experience, which can ultimately lead to their improved perceptions and interpretation of the scores. It is recommended that the test-takers' lived experiences should be taken into

consideration in test use, administration, and provision of feedback in the future. Justice would not only refer to having access to the technical reliability statistics published annually but would also signify their right to benefit from a more transparent system of assessment that the applicants can fully comprehend, communicate with, and trust. This is especially true when an assessment system has such profound effects on the lives of millions of people. In addition to improving test-takers' LAL, test-taking organisations like IELTS need to actively connect with test-takers to reshape their LAL related to knowledge of the local contexts.

Based on our findings, many IELTS candidates have a jagged LAL; that is, they are highly knowledgeable in some aspects of the test (e.g., exam tips, resources, technical grammar terminology, etc.), while a lot of them are confused regarding issues like rating criteria and the levels associated with each band score. Using technical grammar or test terminology by the learners (which are often used by the teachers) could be considered an indicator of negative washback. While it could be argued that students' familiarity with grammatical terms or test-preparation concepts can help them in becoming more autonomous language learners and ultimately improve their general language proficiency, such a tendency is not aligned with the academic/everyday life target language use domain associated with Academic or General Training IELTS (Allen 2016).

We also found that writing was found to be the most challenging skill for the students who were trying to achieve the desired mark. This corroborates the previous findings of Pearson (2019), who reported that 46.2% of instances of poor performance in one or more subtests were related to writing. This is somewhat reflected in the latest IELTS test-taker performance report (2019) by nationality, in which Iranian students received an average of 6.1, 6.4, 6.4 in reading, listening, and speaking of IELTS Academic, respectively, and 5.8 in writing. Further research needs to be conducted to explore the factors that contribute to this issue and how they can be addressed.

This study was, to our knowledge, a pioneering study in employing a data-mining approach to the field of second language assessment. Data mining presents a lot of benefits to second language researchers by enabling them to analyse highly individual data rather than solely relying on analyses at the group or average level (Warschauer et al 2019). Furthermore, it provides access to ecologically valid data collected in genuine educational settings without any intervention. Our study showed that analysing such ample, unmediated data can lead to discovering potentially useful patterns, concerns, and issues which are often undetected in other types of research. In addition, it empowers the researchers to dispense with the traditional labour-intensive steps such as recruiting participants, developing and conducting instruments/interventions like interviews, data entry, etc. Future research in our field can use data mining (either independently or combined

with other approaches) to improve our understanding of the learners, who are increasingly engaged in complex activities in the digital world.

References

Allen, D (2016) Investigating washback to the learner from the IELTS test in the Japanese tertiary context, *Language Testing in Asia* 6 (7), 1–20.

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

Gleason, B (2018) Thinking in hashtags: Exploring teenagers' new literacies practices on Twitter, *Learning, Media and Technology* 43 (2), 165–180.

Hamid, M O, Hardy, I and Reyes, V (2019) Test-takers' perspectives on a global test of English: questions of fairness, justice and validity, *Language Testing in Asia* 9, Article no. 16.

He, W (2013) Examining students' online interaction in a live video streaming environment using data mining and text mining, *Computers in Human Behavior* 29 (1), 90–102.

IELTS (2019) *Test Statistics*, available online: https://www.ielts.org/for-researchers/test-statistics/test-taker-performance

Namugera, F, Wesonga, R and Jehopio, P (2019) Text mining and determinants of sentiments: Twitter social media usage by traditional media houses in Uganda, *Computational Social Networks* 6, Article number 3.

Pearson, W S (2019) 'Remark or retake'? A study of candidate performance in IELTS and perceptions towards test failure, *Language Testing in Asia* 9, Article no. 17.

Saif, S (2021) High-Stakes Test Preparation in Iran: The Interplay of Pedagogy, Test Content, and Context, in Lanteigne, B Coombe, C and Brown, J D (Eds) *Challenges in Language Testing Around the World: Insights for Language Test Users*, Singapore: Springer Nature Singapore, 497–528.

Shih, C M (2007) A new washback model of students' learning, *Canadian Modern Language Review* 64 (1), 135–162.

Warschauer, M, Yim, S, Lee, H and Zheng, B (2019) Recent Contributions of Data Mining to Language Learning Research, *Annual Review of Applied Linguistics* 39, 93–112.

Watson, E (2020) #Education: The potential impact of social media and hashtag ideology on the classroom, *Research in Social Sciences and Technology* 5 (2), 40–56.

5 Analysis of ESL preservice teachers' collaboration with ESL learners through a language assessment development project

Jiyoon Lee
University of Maryland, Baltimore County, USA

In this case study, I document English as a second language (ESL) preservice teachers' perceived learning experience of collaborating with ESL learners during a language assessment development project in the United States. In a semester-long, graduate-level language assessment course, preservice teachers had opportunities to learn ESL learners' perspectives on language assessment through deliberate and focused interaction. The teachers' reflections suggest that the collaboration may provide opportunities to develop their language assessment literacy (LAL). This study's key findings are:

- It is critical to share authentic voices of education participants to enrich their assessment experience
- Trials of assessments in classrooms can help preservice teachers to realize the efficacy of their assessment
- It is necessary to revisit the LAL components to include affective aspects as part of LAL

본 논문은 미국 제2언어 예비교사들이 쓰기 평가를 개발하는 과정에서 학습자들과의 소통을 통해 어떠한 학습 경험을 하였는지를 다룬 사례 연구이다. 한 학기 동안 이루어진 대학원의 언어평가 수업에서 예비교사들은 제2언어 학습자들의 언어 능력, 교육과정 및 평가 경험을 기반으로 쓰기 평가를 설계하고 실제로 평가를 진행하였다. 이 과정을 통해 예비교사들은 자신들의 언어평가에 대한 이해가 향상되었다고 느꼈다. 본 연구는 교육과정 참여자 전체, 특히 학습자와 예비교사의 긴밀한 소통과 실제 경험을 기반으로 평가를 개발해야 한다고 강조한다. 또한 연구자는 언어평가를 더 정확히 이해하고 사용하기 위해서, 언어평가 사용자들의 감정적인 경험도 주요 요소로 다루어야 한다고 주장한다.

Introduction

In this case study, I examined preservice teachers' reflections on collaboration with their ESL learners in an assessment development project in a graduate-level language assessment course. To complete this project, the preservice teachers were required to interact with ESL learners in public schools and design language assessments that were tailored to the needs of their collaborating ESL learners. This study focused on the reflections of four selected preservice teachers. Throughout the project, I was able to observe that the preservice teachers perceived their learners' voices invaluable. I argue that the affective approaches to assessment can be part of language assessment literacy (LAL).

Literature review

While discussions of LAL constructs are still evolving, the general consensus dictates that LAL refers to stakeholders' (1) knowledge of language assessment theories and language learning principles (i.e., assessment knowledge); (2) skills to design, select, and use language assessment (i.e., assessment skills); and (3) understanding of the cultural, pedagogical, political, and social consequences of language assessment (i.e., assessment principles) (Davies 2008, Fulcher 2012, Giraldo 2018, Inbar-Lourie 2008, 2017, Lee 2019). These three factors have guided LAL research to better understand LAL of education participants such as administrators (Deygers and Malone 2019, O'Loughlin 2013), learners (Butler, Peng and Lee 2021), teachers (Vogt and Tsagari 2014), teacher educators (Jeong 2013), and policy makers (Pill and Harding 2013). While earlier LAL research has focused on individual education participants' LAL, later studies started focusing on the collaboration among these education participants (e.g., Baker and Riches 2018, Lee, Butler and Peng 2021). Their studies suggest that the collaboration between the assessment participants may influence each other's LAL and can be part of their LAL. For instance, Baker and Riches (2018) investigated LAL development of teacher-trainers/language assessment specialists and the participating teachers during a series of workshops on language assessment. In their analysis of the participants' interviews, reflections, and samples of teachers' work as evidence of their LAL development, the researchers concluded that the collaboration was conducive to both parties' respective LAL constructions. Lee et al (2021) examined the collaboration among preservice teachers, an in-service teacher, her middle school ESL learners (via the collaborating teacher), and the course instructor as key informants to each other's LAL enhancement. These participants were involved in a development project where preservice teachers designed language assessments for collaborating

ESL learners in real classrooms. An examination of preservice teachers' reflections, the in-service teacher's and course instructor's feedback, and ESL learners' comments about their experience showed that the multi-directional interaction during the project was conducive to each participant group developing their LAL.

In this case study, I focused on the learning experience of preservice teachers, which they identified while interacting with ESL learners. Preservice teachers were engaged in designing an assessment for ESL learners, providing feedback to them and collecting their responses to the feedback. They also reflected upon the collaboration and the learning experience, focusing on their LAL construction. The following research question guided this study: What learning experience do the preservice teachers perceive in the collaboration afforded during a language assessment development project?

Methods

Context

The study was conducted during a 15-week, semester-long, graduate-level language assessment course. The course is offered twice a year (once online and once in person) in the Teaching English to Speakers of Other Languages (TESOL) program at a public university in the United States, mid-Atlantic region. The course introduced (1) language assessment theories (e.g., validity, reliability, authenticity, or fairness); (2) the role of assessment in learning and society; (3) specifics of different types and purposes of language assessments (e.g., listening, speaking, reading, writing assessment, formative, summative assessment, assessment of learning, assessment for learning); and (4) language assessment construction (e.g., assessment specifications, rubric types). It also provided hands-on experience in reviewing, designing, and implementing language assessments. The course was offered in an online, asynchronous format in the semester when the data collection was completed. I was the instructor of the course and provided pre-recorded audio and video lectures and readings from research and practice journals and a textbook. I also held weekly virtual meetings where I addressed students' questions and comments regarding lectures, readings, and assignments. The assignments included (1) discussion postings related to the lecture and readings; (2) language assessment reviews and trials; and (3) language assessment development and implementation for actual ESL learners, which is the target of this study.

Twelve TESOL students attended the course. Their backgrounds, teaching experiences, and career goals were diverse. Some were pursuing a Master's degree in TESOL (comprising 12 to 14 courses), and others were

Table 1 Participants

Preservice teachers	Teaching experience	ESL learners' characteristics	Access to ESL learners
Group 1: Max	• Teaching English as a foreign language (EFL) online • Pursuing a Master's degree in TESOL and ESL certification that permits teaching in US public schools	• Four tenth grade students (15–20 years old) • Low-intermediate to intermediate level in World-Class Instructional Design and Assessment (WIDA) • First language (L1): Spanish	His group member had direct access to ESL learners
Group 2: Samantha	• A practicing high school ESL teacher • Pursuing a Master's degree in TESOL	• Four eleventh and twelfth grade students (16–18 years old) • Intermediate level in WIDA • L1: Spanish, French, Haitian Creole, and other African languages	She had direct access to ESL learners
Group 3: Rachael	• A former speech pathologist • Pursuing a Master's degree in TESOL and ESL certification that permits teaching in US public schools	• Four sixth grade students (11 years old) • Beginning level in WIDA • L1: Burmese, Persian, and Spanish	Her group member had direct access to ESL learners
Group 4: Kelly	• Taught EFL • Pursuing a Master's degree in TESOL and ESL certification that permits teaching in US public schools	• Three fifth grade students (10 years old) • Intermediate to high proficiency level in WIDA • L1: Mandarin or Tamil/ English	Her group worked with a collaborating teacher who had direct access to ESL learners

completing a four-course ESL certificate. Upon graduation, some students planned to obtain US public school teaching qualifications, teach at adult/ refugee centers/private language institutes in the US, or teach English abroad. At the beginning of the semester, the students virtually signed up to make four groups of three people based on their own preferences. They were informed that they would work as a group to develop a writing assessment for actual ESL learners. Each group was required to secure ESL learners for whom they would design the writing assessment. The instructor contacted ESL teachers in local schools to help with those who could not access ESL learners on their own.

Participants

This case study focused on the following four preservice teachers among 12 preservice teachers and their collaborating ESL learners (Table 1).

The preservice teachers were selected for their diverse teaching experience as well as their leading roles in their respective groups. Pseudonyms were used for both preservice teachers and ESL learners.

Language assessment development project

The data for this study were collected from the course's language assessment development project, which provided preservice teachers with opportunities to design and implement language assessments for ESL learners in real classrooms. As course instructor, I decided that the preservice teachers would design a writing assessment so they could observe ESL students' performances directly (i.e., through performance assessment). In addition, the writing assessment was logistically less complicated to implement than speaking assessment. The project consisted of the following stages: (1) planning; (2) assessment design; (3) implementation; (4) evaluation and feedback; and (5) reflection.

At the planning stage, the preservice teachers submitted two written reports: the test-taker report and the construct report. These submissions documented their ESL learners' general L2 proficiency based on the WIDA (2020), one of the widely used standard-based language assessments for public school ESL learners in the US. They also collected information about the learners' first language proficiency/literacy levels, length of time studying English at a US school, and other relevant information. Then, the preservice teachers reviewed the curriculum and lesson goals/contents. Based on that information, they defined the writing skills to be assessed (e.g., vocabulary and grammar knowledge, paragraph writing, organization). Figure 1 shows Rachael's group constructs example. To select the constructs, her group adapted one of the Common Core State Standards (NGA Center 2010) for their assessment, which is a set of widely adopted centralized curriculum standards in the area where this data was collected. They selected 'organization/format', 'content', 'grammar', and 'convention/mechanics' as their constructs.

At the design stage, each group was free to design a formative or summative writing assessment for online or paper-and-pencil delivery. Their collaborating teachers' or students' needs were the determining factors for their decisions. For instance, Kelly's group decided to design a summative writing assessment to assess what their collaborating learners learned during the first quarter. Based on the test-taker characteristics as well as the constructs for the writing assessment, the preservice teachers decided on the specifications of the assessment and developed writing prompts and rubrics. Then, the groups implemented their assessment with their collaborating ESL students.

At the evaluation stage, the preservice teachers completed (1) individual grading of their ESL learners' writing performances with their rubrics; (2) discussions of their individual grades to come to a common grade; and

Figure 1 Constructs example

Constructs

1. Organization/Format
    a. Sub-construct: Identify and/or produce the five components of a letter: date, greeting, body, closing, signature.
    b. Sub-construct: Produce a minimum of four (4) sentences in a structured task.
2. Content
    a. Sub-construct: Produce a topic sentence to state what they are thankful for.
    b. Sub-construct: Produce at least two supporting sentences to describe what the person did.
    c. Sub-construct: Produce at least one supporting sentence to state why they are thankful.
3. Grammer
    a. Sub-construct: Identify and/or produce sentences with appropriate structure/word order.
4. Conventions/Mechanics:
    a. Sub-construct: Identify and/or produce appropriate capitalization at the beginning of a sentence and/or for proper nouns.
    b. Sub-construct: Identify and/or produce ending punctuation in a sentence.

(3) written feedback for their learners. Individual preservice teachers also provided written feedback to each student they worked with.

After receiving the preservice teachers' feedback on their performance, the learners were also asked to answer the following prompts:

1. What do you think of the writing test you took? Do you think the test showed your English writing ability?
2. What did you like or dislike about the writing test?
3. How do you think the preservice teachers can improve their test?
4. Did you find the preservice teachers' feedback helpful for you to improve your writing?

Depending on the preservice teacher groups' contexts, they directly interviewed or surveyed learners (Max, Samantha and Rachael) or received help from their collaborating teachers (Kelly). To answer these questions, the students responded to these prompts in writing or orally.

At the reflection stage, the preservice teachers critiqued their own groups' assessments and reflected on their test-takers' reaction to their feedback. The preservice teachers were required to address the following prompts in their written reflections.

1. What strengths and weaknesses did you find in your writing assessment?
2. Do you think your group's writing assessment elicited students' writing ability?

3. Do you think your feedback was effective for the students who took your group's assessment? How do you know?
4. How can you improve your writing assessment?
5. What did you learn from the interaction with the ESL learners?

Data and data analysis

Table 2 presents the data used for this study. The data includes (1) the four preservice teachers' written reflections on the learning experience in relation to LAL construction; and (2) the learners' perceptions of the assessments that the preservice teachers designed for them. The preservice teachers submitted their learners' responses to the prompts as part of their assignment.

Table 2 Data

Participants	Contents	Formats
Preservice teachers	A final paper for the course with reflections on the learning experience during the assessment development project	Written
ESL learners	Responses to questions to reflect on their participation in the new assessment	Written or oral

Based on the three components of LAL provided in the earlier section (i.e., assessment knowledge, assessment skills, and assessment principles), these reflections were then analyzed deductively. Idea units were used for analysis. An idea unit is a chunk of cohesive information in written or spoken discourse (Kroll 1977). Following Nowell, Norris, White and Moules (2017) and Saldaña (2013), I conducted double data coding.

Results and discussions

This research intends to document how the preservice teachers perceived the collaboration with ESL learners and what learning experiences they identified for themselves. The findings suggest that the preservice teachers recognized the importance of implementing their assessment to actual learners and hearing ESL learners' authentic voices. Their reflections hinted that they utilized the ESL learners' voices to develop their LAL.

Preservice teachers' perceived learning experience

The data suggested that the preservice teachers perceived progress in their LAL development in all three LAL components through collaboration

with the learners. In terms of assessment knowledge, the preservice teachers reported that they became aware of learners' needs and characteristics as the starting point of their assessment design. Kelly stated that the language assessment development project and the interaction with the learners helped her understand the importance of 'placing learners' cognitive and linguistic needs in the center stage' in developing the language assessment. She also reported, '[W]hile listening to their [the ESL learners'] feedback on the assessment experience, I learned a valuable lesson: individual test-takers' characteristics matter and must be prioritized in all stages of the assessment development'. She also noted that hearing ESL learners' responses to their assessment helped her 'self-reflect' and 'grow as an educator'. Max addressed the importance of the data that his group's assessment provided. Observing the relationship between the learners' performance and his collaborating teacher's instruction, Max described how he connected the teacher's instructional goals, his group's target constructs, and the learners' performance, and provided tailored feedback to his learners.

The preservice teachers' reflections presented how they ensured their assessment quality while sharing learners' positive responses to their assessment. Although the learners did not use any assessment-related jargon, it seemed that the preservice teachers associated their learners' feelings with assessment quality. One of Rachael's learners' comments is shown in Figure 2 below. This student viewed the assessment experience as fun or positive because she felt that she could write all she knew and wanted. In other words, the student felt that the assessment measured what she could do in writing (i.e., a validity issue). This comment from an ESL learner helped the preservice teachers realize the possible connections between learners' motivations and positive orientations toward theoretically sound (valid) assessments.

The preservice teachers acknowledged that as their assessment was designed based on collaborating ESL learners' needs, interests, and proficiency levels, the students were fairly likely to have positive reactions to the assessments.

The preservice teachers also stated that the collaboration with the ESL learners helped them realize what language assessment skills they have gained and what more was needed. In her reflection, Samantha agreed with her students' comment about clarity of instructions, reflecting that 'the instructions for all three sections could be simpler and clearer, so the test-takers know exactly what they need to do'. She also reported that the ESL learners' performance helped her group realize the efficacy of their rubric. She explained that due to ambiguous descriptors in their rubric, her evaluation of the ESL learners' performance was not sufficiently precise, and her feedback was less constructive than she had hoped.

As seen in Figure 3 below, one of Kelly's ESL learners provided informative suggestions related to the assessment format – more space for

Figure 2 Student's response 1

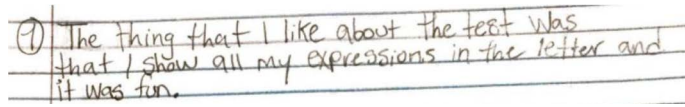
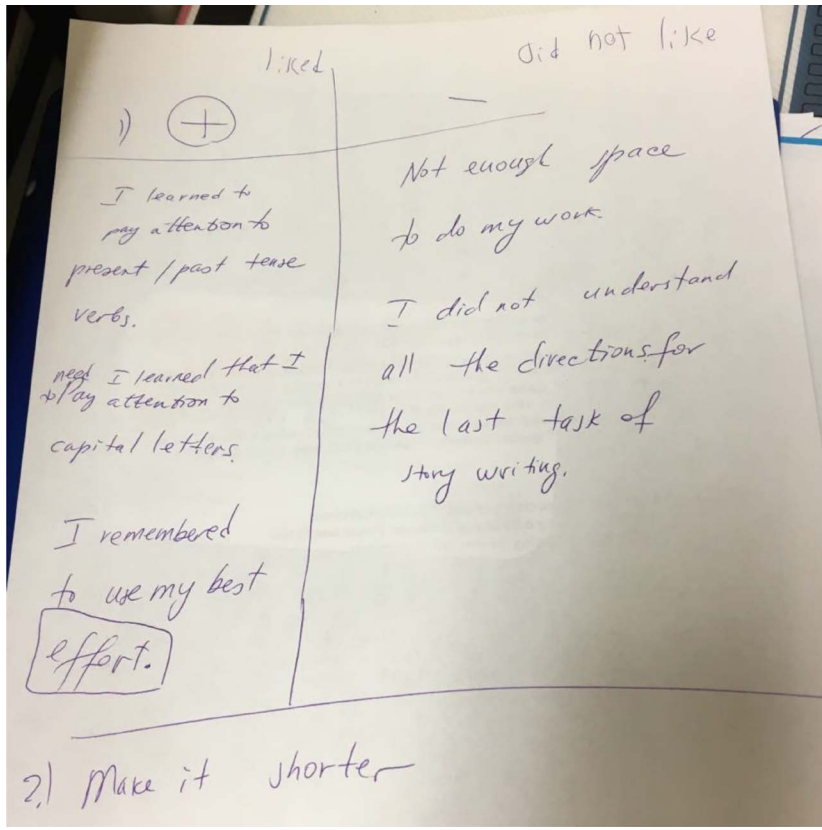


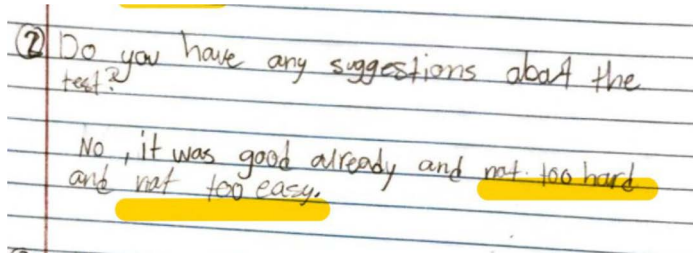
Figure 3 Student's response 2



writing and providing clear directions for the testing items. In her reflection, Kelly noted that one of the weaknesses of the assessment was the insufficient space for writing.

As seen in Figure 4 below, multiple ESL learners, including one of Rachael's, talked about the difficulty level of the assessment and their performance. Rachael addressed the positive comments regarding learners' perception of the difficulty level when assuring the quality of their assessments.

Figure 4 Student's response 3



The discussions of assessment principles centered on the feedback that the preservice teachers provided to the learners. The learners' overall positive perceptions of the assessment were reflected in their reaction to the teachers' individualized feedback. Max shared his collaborating students' comments about the integrity of the feedback they had received (i.e., 'feedback help me because he is giving me a lot of advice') and ways he could use it (e.g., 'that help me for Keep working hard'). Kelly also reported that her collaborating learners addressed the feedback she provided to them as a positive experience when they were asked what they liked about the newly designed language assessment (see Figure 3). Samantha described her ESL learners' inquiries about the preservice teachers' feedback and ways to get 'good grades' to respond to the feedback they had received.

As noted, the learners indicated their receptiveness to the preservice teachers' feedback and planned to take it up. Nonetheless, the data is limited in that the preservice teachers could not observe how the learners actually used their feedback in their language learning once the language assessment development project was completed. To maximize the opportunity for preservice teachers to experience the efficacy of their feedback, the language assessment development project could include one more stage. In that stage, the learners would be asked to revise their work based on the feedback they received, and the preservice teachers would evaluate their learners' revisions and the effectiveness of their feedback.

LAL components suggested: Affective approaches to assessment

Although education participants' affective approaches to assessments are not included in the current LAL components, the preservice teachers noted that learners' perspectives on language assessment and teacher engagement may include consideration of their feelings. Max mentioned that the way the learners felt about their assessment could impact their performance on assessment. Reporting the impression he got from the learners' comments

about their assessment, he wondered about ways to promote learners' willingness to participate in learning by providing less stressful, safe, and fair assessments. Kelly also stated, 'when designing the assessment, my primary objective was to provide them with a valuable, enjoyable experience and support their learning'. She shared how her feedback was received by one of her learners, stating 'he reevaluated his responses and suggested (a) the story's main idea for the first paragraph and (b) an additional supporting detail' in his plan for future writing. Samantha also observed that when she clearly explained the purposes of the assessment (i.e., formative and helping them improve their writing), her students showed more interest and were engaged in the assessment they designed.

It is probable that the prompts asked to elicit the ESL learners' perspectives on the assessment focused on their affective approaches to assessment (e.g., asking about their likings). However, the comments made by the students regarding the validity of the test and other assessment quality indicate that if the ESL learners are properly guided with more precise and concrete prompts, we may hear learners' perspectives on assessment in greater detail, including suggestions for how to enhance tests and support learners in using test results for their own learning.

Conclusion

The findings corroborated the importance of connecting education participants closely and having their voices heard through assessment trials. Preservice teachers' reflections confirm that they perceived the interaction with the ESL learners as a valuable experience to evaluate the efficacy of their assessment and to gauge their receptiveness to it. As one of the preservice teachers mentioned, learners could provide enriching information about their assessment experience. This study encourages additional discussions on how to involve learners in teacher education and encourage them to share their perspectives on their assessment experiences. Despite the limited data, the study also makes the case for revisiting the concepts of LAL and initiating discussions of integrating affective components in LAL, particularly, learners' LAL. The language assessment development project is worth replicating in a range of educational contexts (e.g., a semester-long course or a day workshop) and interface modes (e.g., online or face-to-face) even though it necessitates careful logistical management by teacher educators.

Acknowledgements

I have great admiration for hardworking preservice teachers who strive to learn effective teaching and assessment strategies to better support our

rapidly expanding ESOL community. I am especially grateful to the four preservice teachers who took part in this study and to the ESOL students who kindly completed the newly developed assessment and shared their feedback.

References

Baker, B A and Riches, C (2018) The development of EFL examinations in Haiti: Collaboration and language assessment literacy development, *Language Testing* 35 (4), 557–581.

Butler, Y G, Peng, X and Lee, J (2021) Young learners' voices: Towards a learner-centered approach to understanding language assessment literacy, *Language Testing* 38 (3), 429–455.

Davies, A (2008) Textbook trends in teaching language testing, *Language Testing* 25 (3), 327–347.

Deygers, B and Malone, M (2019) Language assessment literacy in university admission policies, or the dialogue that isn't, *Language Testing* 36 (3), 347–368.

Fulcher, G (2012) Assessment literacy for the language classroom, *Language Assessment Quarterly* 9 (2), 113–132.

Giraldo, F (2018) Language assessment literacy: Implications for language teachers, *Profile: Issues in Teachers' Professional Development* 20 (1), 179–195.

Inbar-Lourie, O (2008) Constructing an assessment knowledge base: A focus on language assessment courses, *Language Testing* 25 (3), 385–402.

Inbar-Lourie, O (2017) Language assessment literacy, in Shohamy, E, May, S and Or, I (Eds) *Language Testing and Assessment* (Third edition), Cham: Springer, 257–268.

Jeong, H (2013) Defining assessment literacy: Is it different for language testers and non-language testers?, *Language Testing* 30, 345–362.

Kroll, B (1977) Combining ideas in written and spoken English: a look at subordination and coordination, in Ochs Keenan, E and Bennett, T L (Eds) *Discourse Across Time and Space*, Southern California Occasional Papers in Linguistics 5, Los Angeles: Department of Linguistics, University of Southern California, 69–108.

Lee, J (2019) A training project to develop teachers' assessment literacy, in White, E and Delaney, T (Eds) *Handbook of Research on Assessment Literacy and Teacher-Made Testing in the Language Classroom*, Hershey: IGI Global, 58–80.

Lee, J, Butler, Y G and Peng, X (2021) Enhancing preservice teachers' language assessment literacy through multiple stakeholder involvement, *Languages* 6, 213.

National Governors Association Center for Best Practices and Council of Chief State School Officers (2010) *Common Core State Standards*, Washington, DC: National Governors Association Center for Best Practices, Council of Chief State School Officers.

Nowell, L S, Norris, J M, White, D E and Moules, N J (2017) Thematic analysis: Striving to meet the trustworthiness criteria*, International Journal of Qualitative Methods* 16, 1–13.

O'Loughlin, K (2013) Developing the assessment literacy of university proficiency test users, *Language Testing* 30 (3), 363–380.

Pill, J and Harding, L (2013) Defining the language assessment literacy gap: Evidence from a parliamentary inquiry, *Language Testing* 30 (3), 381–402.
Saldaña, J (2013) *The Coding Manual for Qualitative Researchers*, London: Sage Publications.
Vogt, K and Tsagari, D (2014) Assessment literacy of foreign language teachers: Findings of a European study, *Language Assessment Quarterly* 11 (4), 374–402.
WIDA (2020) *WIDA English Language Development Standards Framework, 2020 Edition Kindergarten—Grade 12*, Madison: Board of Regents of the University of Wisconsin System.

6 Assessing spontaneous oral language use in instructed contexts

Kathryn Macfarlane, Director, Reading's a Breeze! Ltd

This case study presents the findings of a professional learning (PL) program evaluation study, showing how developing language assessment literacy (LAL) for both teachers and learners led to transformational change in oral language learning outcomes. It includes:

- The context of the language education landscape in Australia
- A brief description of the Autonomous Language Learners (ALL) Approach, which was the foundation of the PL program
- Relevant details regarding the PL program and the changes schools implemented as a result of their participation
- A case study of the school in which transformational change in learning outcomes was most strongly demonstrated

Context

Languages have been part of Australia's earliest history and continue to be a part of the educational landscape; over 200 different languages are spoken – and dozens of these taught – across what is officially a monolingual, English-speaking nation. However, there have been systemic failings in language acquisition in school-based programs (Liddicoat et al 2007). The state of Victoria has arguably led the field; Languages is a mandatory key learning area with school registration being contingent upon offering a Languages program (Victorian Registration and Qualifications Authority 2019). The Department of Education and Training (DET) defines quality language programs, with a recommended allocation of (as a minimum) 150 minutes spread evenly across the week (DET 2019a). The Victorian Curriculum – Languages and its Achievement Standards were drafted based on this aspirational recommendation. However, the average primary years language program structure in Victoria has remained a single weekly lesson of less than 60 minutes since annual survey reporting began decades ago (DET 2019b).

This is possibly because within the constraints of minimum mandatory requirements, principals have autonomy over budgeting, timetabling and curriculum implementation within schools.

The key expectation of language learning is developing the ability to speak; however, Gonzalez Humanez and Arias Rios (2009:2) identify that 'oral interaction is one of the most difficult competences to develop', and that 'oral interaction skills are often neglected in classroom environments'. Various reports have identified this lack of oral language development in Australian primary years language programs (Education Department of Western Australia and the Commonwealth Department of Education, Training and Youth Affairs 1998, Lo Bianco 2009, Scarino et al 2011).

Despite these numerous published findings of learning outcomes not meeting expected standards, reporting to parents has lacked transparency. Language teachers must report using a 5-point scale, often with no opportunity for differentiated commentary. Many school principals instruct language teachers to report that the majority of learners meet the expected standard, regardless of the actual learning outcomes.

Assessment of oral language in classroom contexts is also problematic. When it takes place at all, it focusses on performed role-plays and rehearsed speeches, neither of which reflect an ability to interact in, or spontaneously use, the target language. The disconnect between assessment practices, reporting to parents and actual student achievement has led to poor public perceptions and language education in Australia being described as 'in crisis' (Group of Eight 2007). To improve language education in Australian schools, transformational change is needed at multiple levels, including in the area of language assessment literacy (LAL) for school leaders, teachers, and learners themselves.

The Autonomous Language Learners Approach

In response to this need for transformational change to improve language learning outcomes (in particular the use of oral language for classroom interaction), the Autonomous Language Learners (ALL) Approach was developed (Macfarlane 2020). It is a multi-disciplinary combination of eight key strategies, derived from the fields of Organisational Change Management, Applied Linguistics and Learner Autonomy. Table 1 lists these eight key strategies and their purpose.

The ALL Approach was the foundation of a PL program co-designed and co-delivered with Catholic Education Melbourne (CEM) to 215 teachers representing 89 schools (it was strongly recommended that schools nominate a team to attend, including as a minimum the language teacher, the literacy coordinator and the principal or senior leader). The 18-month-long program spanned three school years for each cohort. Content was delivered via a

combination of face-to-face workshops, online webinars, and in-school coaching visits by CEM language experts.

Table 1 The eight key strategies of the ALL Approach

Field	Strategy	Purpose
Organisational Change Management	#1 – Develop teachers' language assessment literacy for spontaneous oral language use	Gather baseline data for evidence of the need for change; measure the subsequent impacts of change on learning outcomes
	#2 – Secure the support of school leadership	Remove barriers to change
Applied Linguistics	#3 – Increase frequency of weekly contact with the target language	Enable retention of new vocabulary and language structures introduced
	#4 – Focus lesson content on high-frequency vocabulary and language structures necessary for classroom interaction	Increase opportunities for language use by facilitating classroom communication in the target language
	#5 – Use intentional teaching gestures (Maxwell 2001, Wilks-Smith 2019)	Support comprehension, acquisition and retention of new vocabulary
Learner Autonomy (Little 2009)	#6 – Support learners to set and revise SMART personal learning goals (Doran 1981, Madden 1997)	Set clear expectations relating to desired outcomes, engage learners in learning process
	#7 – Provide learners with tools with which to self-monitor their own progress	Develop learners' language assessment literacy, create goal–action–result feedback loop
	#8 – Revise reporting process	Focus on goals of oral language development, use learner self-assessments to inform results reported, allow learners to draft their own language learning reflection

It was determined that during this program, Strategies #1, #6, #7, and #8 all required significant development in LAL on the part of both teachers and learners. Strategy #1 was the first to be addressed with teachers in Workshop #1 during the first trimester of the PL program, before changes were introduced to their school's Languages program. The notion of assessing spontaneous language use rather than rehearsed production was challenging for many teachers, as they had never attempted to do so. In addition, the logistics of capturing spontaneous language use for a class of students in a single weekly lesson, then reviewing and assessing that data, was seen as prohibitive.

The initial purpose of focussing on Strategy #1 was to gather baseline data as evidence of need in this area, in order to secure leadership support

for the necessary changes. The self-assessment task that students would subsequently use was introduced at this point. The task needed to be familiar to young learners, and able to be administered *en-masse* to a whole class within a limited period due to the timetabling constraints outlined above, as well as being quantifiable and relevant to the spontaneous use of oral language. CEM commissioned the development of an online speech-to-text tool ('SpeakUp!') designed to capture student-dictated responses to a picture narration task.

SpeakUp! captures information on characteristics of the spoken output. The following measures were selected for our purposes: the number of word types (NWT), longest sentence length (LSL) and average sentence length (ASL). Subsequently, number of words per minute was also introduced as a measure of spoken fluency. When students log into SpeakUp!, they see their previous performance and set new goals for the current task (see Figure 1). They are then presented with a series of three or more prompt images during 2 minutes' planning time, followed by 7 minutes in which they record their response. Their performance against the target measures is displayed and updates in real time as they dictate. With the constraints of voice-to-text technology, the transcription is never 100% accurate; after completing the task, students are able to edit their transcription to correct any errors they are able to identify (see Figure 2). After completing their review, students are asked to reflect on their performance and select an area of focus for improvement in their next task. A learning tip is displayed, relevant to the area of focus they have selected.

Following the introduction of SpeakUp! at the initial workshop, CEM coaches visited all 89 schools to assist teachers with the implementation of this web application with their classes. The results were discussed during a webinar. Across all schools, it was found that the number of word types was limited to a maximum of 50 words, regardless of the number of years of language study. The majority of these were topic-based (numbers, colours, food) and unable to be used to form sentences without additional vocabulary (i.e. verbs, conjunctions, prepositions) which had not been acquired. Teachers reported anecdotally that spoken interaction in the classroom was either absent or limited to a small number of formulaic greetings and questions (hello, how are you, can I go to the toilet?). Many teachers were shocked at the low number of words their students were able to spontaneously produce for the task, and their inability to respond to the images depicting common classroom situations. This was an eye-opening experience, highlighting for teachers that their current assessment practices had not captured their students' lack of vocabulary development.

With these results in hand, all teachers were motivated to introduce changes. A focus on high-frequency vocabulary and language structures for classroom communication was strongly adopted in all participating schools

Figure 1 Goal setting in SpeakUp![1]

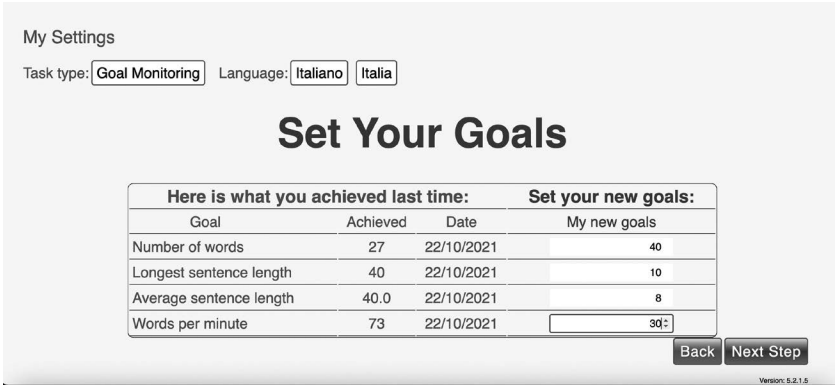


Figure 2 Editing the Narrated Response

(Strategy #4), as was the use of intentional teaching gestures (ITG) to support comprehension and retention (Strategy #5). Leadership support (Strategy #2) was secured more strongly in some schools than others, resulting in varied increases in frequency of weekly contact (Strategy #3). Where this increased frequency did occur, it was achieved through generalist classroom

1 All images are copyright Catholic Education Commission of Victoria.

teachers agreeing to become co-learners of the target language, using acquired vocabulary and language structures throughout the day, every day, and modelling effective learning behaviours for their students. They were supported in their learning by a language specialist (either a teacher or an assistant). These generalist teachers participated in the ALL PL program, alongside their specialist language colleagues.

In addition to the quantitative data gathered using SpeakUp!, discussions during subsequent face-to-face workshops helped teachers to further develop their repertoire of tools for qualitative assessment of spontaneous use of oral language for classroom communication. Already familiar with the use of rubrics for student self-assessment in other learning areas, the generalist teachers pioneered the use of these with their students for language. They developed initial drafts of the rubrics during PL program workshops, which they then refined in a co-creation process with their students at school (see Table 2). In a peer-learning process, this idea spread to the specialist language teachers in each cohort of the ALL PL Program.

A peer observation template was also developed by workshop participants, allowing classroom learning buddies to rate each other's language use during small group activities, as well as a vocabulary audit sheet, allowing learners to colour-code words they were confident using (green), needed more practice using (orange), or couldn't yet use (white) from a list of identified high-frequency vocabulary.

The ALL Approach PL Program Evaluation Study

The evaluation study investigated outcomes of the ALL PL Program at two levels: i) the changes made to the school's language program, and ii) the resulting impacts on student oral language acquisition and use. Four overlapping cohorts were involved in the study between 2017 and 2020. Student achievement data were gathered by teachers in schools as part of their revised assessment practices, and 17 schools provided these data for analysis. Pawson's (2002) Realist Evaluation Model was used to evaluate the impacts of the ALL PL Program on the two levels described above. In addition to the measures identified for student self-assessment, the researcher also evaluated the number of key points (NKP) included in the picture narration task performed in SpeakUp!.

As seen in Figure 3, across all four cohorts, increased vocabulary acquisition only occurred where school leaders agreed to increase the frequency of weekly contact with the target language (Schools G–N). This data demonstrates the pre-requisite nature of leadership support and frequency of contact for successful language learning outcomes.

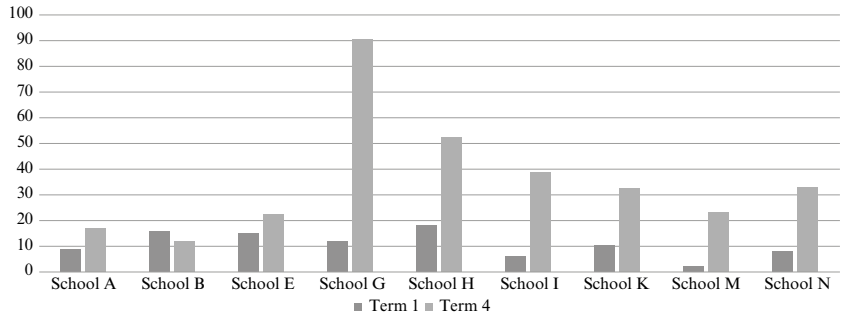
However, increased frequency alone did not account for the most substantial transformation in learning outcomes, nor did the focus on

Table 2 Student language use rubric

	Being Proactive	Making Meaning	Communicating	Language for a Purpose
Brand New	I am trying to learn and listen to other people speaking the language	I join in when we practise words and sentences		
Just Starting	I speak the language when someone reminds me	I can use single words and some memorised phrases to answer questions	I can ask some questions that I've memorised such as "Can I go to the toilet?" and "How are you?"	I can label some things on a picture of a classroom
Gaining Confidence	I am beginning to speak the language without being reminded	I can put some chunks together to make new sentences (e.g. [Can I go] to the office?, [Can I go] outside?	I can contribute 1 or 2 turns in an unscripted conversation	I can describe 1 or 2 things about a picture of a classroom with some words (e.g. "teacher angry boy stop")
Showing Progress	I often try to use the language before/rather than English in class	I can use the vocabulary I know to make quite a lot of different questions and sentences	I can contribute 3 or 4 turns to an unscripted conversation and check understanding with my partner	I can describe 3 or 4 things about a picture of a classroom using words and sentences
Look at Me Now!	I always try to use the language rather than English in class	I can say most things I need to in the classroom using the language	I can keep a conversation going and find different ways of saying things to make sure my partner and I understand each other	I can make up a story based on a picture of a classroom using more than 5 sentences

high-frequency vocabulary and use of ITGs (both of these strategies were strongly adopted across all schools). While students in School G showed the strongest growth in vocabulary acquisition, students in School H showed the strongest overall improvement across all measures (Number of Word Types, Average Sentence Length, Number of Key Points, and the Language Use Rubric). This is the school which implemented Strategies #6 (goal setting) and #7 (self-monitoring of progress) most thoroughly and consistently; it is therefore worth considering in closer detail.

Figure 3 Median increase in scores for NWT during first year of change implementation[2]



Case study: Our Lady of Collaboration Primary School[3]

Among the population of just under 300 students at Our Lady of Collaboration Primary School, 13% of families report using a language other than English at home (Australian Curriculum Assessment and Reporting Authority 2020); it is an essentially monolingual community in comparison to the multicultural and multilingual communities of many of Melbourne's other schools. In 2015, the Italian teacher retired and Alison (the school principal) was eager to use the opportunity to reimagine the language program.

In 2016, with support from CEM, Alison began planning for change. This leadership-driven initiative and the collaborative manner in which it was implemented were two of the factors which predisposed this school community to successful implementation of the ALL Approach. Initially, the Italian teacher was not replaced. Alison and her leadership team wanted to take 12 months to carefully plan the future of their language program. In order to continue meeting the mandatory requirement for provision of a language program, all classroom teachers were asked to include a language of their choice in their daily lesson planning.

Concurrently, three classroom teachers enrolled in Cohort 1 of the ALL PL Program. A Languages Leadership Team was formed (including these three staff, the school principal and deputy principal) and attended in-school coaching sessions. French was selected as the new target language, as no

2 Of the 17 participating schools, not all provided both baseline data and post-implementation data for comparison.
3 Names of the school, staff and suburb are pseudonyms.

teacher had any prior knowledge (so no class would be at an advantage). It was emphasised that classroom teachers would become *co-learners*, not teachers, of French. Their role would be to model effective learning strategies for their students. A position for a native-speaking French Assistant was created to support the classroom teachers in their learning and assist with planning. This staff member became an integral part of the French Leadership Team. The school was actively involved in Cohorts 1, 2 and 4 of the ALL PL Program, demonstrating ongoing commitment to the improvement of French learning outcomes.

Changes to the language program after participation in the ALL PL Program

Teachers (as co-learners) and students at Our Lady of Collaboration participated in gathering baseline data and immediately began setting learning goals, which were regularly revised (Strategy #6). Initially, their focus was on increasing their active range of high-frequency, functional vocabulary. Once all learners had reached close to (or exceeded) 100 words, the focus shifted to increasing sentence length, while continuing to expand the vocabulary repertoire. In addition to these quantifiable measures from the picture narration task, learners set themselves qualitative goals for their use of French throughout the day. Learning goals were a regular feature of classroom discussion and were revised at least twice per term.

Our Lady of Collaboration has made regular use of Speak Up! since 2017 (Strategy #7), representing the most significant use of the web application across all participating schools. Each class from Year 3 to 6 completes various practice activities and at least one goal-monitoring task per term.

In addition to Speak Up!, templates for peer observation and personal goal-setting for use of French language during small group activities were introduced. During group activities, one student is allocated the role of observer, keeping a tally of the language use of their buddy during the activity. The participating student reviews these observations at the end of the activity and sets goals for next time. The vocabulary audit template is also used regularly, comprised of all the words introduced at the school to that point.

Through this process of goal setting and progress monitoring, both student-learners and teacher-learners have become more aware of the need to break a task (using oral language for classroom communication) into measurable components (vocabulary range, sentence structure awareness, autonomous language-use behaviours); to set goals for each of those components; to monitor progress, and revise goals according to the results.

Where results were not as desired (high NWT but low ASL and NKP), goals were revised to focus on the less developed areas. New goals led to new, more productive language-use behaviours. Teachers and students were thus actively engaged in developing their LAL, performing formative assessment activities in monitoring progress against their goals. The link between goals, assessment practices, revised behaviours and resulting learning outcomes became clear to them, creating a spiralling cycle of success which Alison reported as contributing to the overall learning culture of the school.

In 2017, reports were still drafted by classroom teachers, but individualised comments included the language goals students had set themselves and how they had progressed towards achieving them. In addition, students chose which of their work samples were shared with parents in order to best illustrate their progress towards their personal goals.

Although the official reporting process continued to be teacher-driven, use of French for classroom interaction was clearly identified as a key indicator of progress. Students were aware of this and had input via their self-assessments (Strategy #8). In early 2020, the Languages leadership team expressed a desire to explore the logistical and software constraints of having students draft their own official reports to involve them even more in the assessment process. This was a work in progress at the time the evaluation study concluded.

Student language learning outcomes

By October 2017 (nine months after changes were first introduced), school visits provided evidence of the potential of this new model. Classroom teachers who had no prior knowledge of French were integrating it into their daily communication with students. The children were responding in kind. Comments by staff and observations by CEM coaches confirmed the learning culture of the school had changed; French had become 'normalised' and was something students were curious about. Multilingualism (and by association multiculturalism) had gained prominence as a key feature of the school's identity. Classroom teachers' utterances were not assumed to be a correct model of French language (this was the role of the language assistant, Amélie); rather, it was their learning behaviours which were emulated by students.

In the early stages, responses to the picture narration task consisted of a 'language dump', in which students dictated as many words as they could think of, regardless of their relevance to the task, in order to achieve the highest NWT score. Later, students began constructing more meaningful responses. For example, students quickly put into practice a strategy of using '*et*' [and], or using multiple adjectives to lengthen their sentences: *'Est-ce que tu aimes chanter et que tu aimes danser est ce que tu aimes jouer*

de la musique?' [Do you like singing and do you like dancing and do you like playing music?].[4] They were developing both their awareness of language assessment and strategies to ensure successful results against the stated criteria.

In order to fully appreciate the progress made, it is best to take a deeper look at the language samples of an individual student. Angela commenced studying French in 2017, in Year 3. She had no prior knowledge of the language and this was reflected in her baseline language sample; she was only able to use the formulaic greetings she had learnt during the first week. During Angela's last sample taken in Term 4 2019, she was able to produce a structured response using 74 unique words with an average sentence length of 6.0. Her response included 10 key points, all directly related to the image prompts, and her speech rate was 44 words per minute (see the Appendix for the full transcript and translation).

Discussion

The 'teachers as co-learners of language' model adopted at Our Lady of Collaboration enables daily use of the language (Strategy #3); a pre-requisite for cumulative language acquisition. This alone is worth noting for policy and program planning purposes as it provides a resolution to the 'intractable problem' of teaching conditions in Australian Languages education described by Liddicoat et al (2007:118). However, it was the adoption of a strong focus on improving LAL (Strategy #1), including goal-setting (Strategy #6) and self-assessment (Strategy #7), which led to the greatest increase in learning outcomes in comparison to other participating schools. Revised reporting practices (Strategy #8) were also adopted more strongly at this school, although it continued to be a work in progress. The increased LAL for both students and teachers led to an understanding of *how* to assess the problematic area of spontaneous oral language production, and which *success criteria* were needed to generate the desired changes in learning and language use behaviours.

The language sample cited above and provided in the Appendix is reflective of the language used across the school; it incontestably contains numerous errors. However, these 'errors' constitute evidence of spontaneous, creative language use, a growing vocabulary range, and an emerging understanding of sentence structures; all identified as the desired outcomes.

In addition to the emergence of a new and promising model for language education, the results of this program evaluation study highlight the need for LAL development, to ensure assessment is linked to the skills we most

4 Year 6 student, December 2017, Average Sentence Length 4.8 in this overall sample.

want students to develop. Further study of the results at this school after a complete primary years cycle of seven years will undoubtedly provide additional insights.

References

Australian Curriculum Assessment and Reporting Authority (2020) *My School*, available online: /myschool.edu.au/

Council of Europe (2020) *CEFR Self-Assessment Grid*, available online: rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=090000168045bb52

Department of Education and Training (2019a) *What Makes a Quality Language Program?*, available online: www2.education.vic.gov.au/pal/languages-education/guidance/what-makes-quality-language-program

Department of Education and Training (2019b) *Languages Provision in Victorian Government Schools, 2018,* available online: www.education.vic.gov.au/Documents/school/teachers/teachingresources/discipline/languages/eduState-languages-provision-report-2018.pdf

Doran, G (1981) There's a S.M.A.R.T. way to write management's goals and objectives, *Management Review* 70 (11), 35–36.

Education Department of Western Australia and Commonwealth Department of Education, Training and Youth Affairs (1998) *Factors influencing the uptake of Modern Standard Chinese, Korean, Modern Greek and German at primary and secondary level in Australian schools*, available online: academia.edu/16441681/Factors_influencing_the_uptake_of_Modern_Standard_Chinese_Korean_Modern_Greek_and_German_at_primary_and_secondary_level_in_Australian_schools

Gonzalez Humanez, L E and Arias Rios, N R (2009) Enhancing oral interaction in English as a foreign language through task-based learning activities, *Latin American Journal of Content & Language Integrated Learning* 2 (2), 1–9.

Group of Eight (2007) Languages in crisis: A rescue plan for Australia, available online: www.lcnau.org/pdfs/Go8_Languages_in_Crisis_Discussion_Paper.pdf go8statements/2007/Go8_Languages_in_Crisis_Discussion_Paper.pdf

Liddicoat, A J, Scarino, A, Curnow, T J, Kohler, M, Scrimgeour, A and Morgan, A-M (2007) *An Investigation of the State and Nature of Languages in Australian Schools,* Department of Education, Employment and Workplace Relations, Commonwealth of Australia.

Little, D (2009) Language learner autonomy and the European Language Portfolio: Two L2 English examples, *Language Teaching* 42, 222–233.

Lo Bianco, J (2009) *Second Languages and Australian Schooling*, ACER Press.

Macfarlane, K (2020) *Transformational Change for Primary Years' Foreign Language Programs – Developing Oral Language Skills for Spoken Interaction in the Classroom*, Thesis, Monash University, available online: doi.org/10.26180/13140449.v1

Madden, L E (1997) Motivating students to learn better through own goal-setting, *Education* 117 (3), 411–414.

Maxwell, W (2001) *Evaluating the effectiveness of the Accelerative Integrated Method for teaching French as a second language*, Master's dissertation, University of Toronto.

Pawson, R (2002) Evidence-based policy: In search of a method, *Evaluation* 8 (2), 157–181.

Scarino, A, Elder, C, Iwashita, N, Kim, S H O, Kohler, M and Scrimgeour, A (2011) *Student Achievements in Asian Languages Education – Part 2: Descriptions of Student Achievement*, available online: docplayer.net/20508856-Student-achievement-in-asian-languages-education-part-2-descriptions-of-student-achievement.html

Victorian Registration and Qualifications Authority (2019) *Guidelines to the Minimum Standards for School Registration*, available online: www.vrqa.vic.gov.au/registration/Pages/schminsdards.aspx

Wilks-Smith, N (2019) *Learning with Intentional Teaching Gestures: Japanese foreign language output in the primary* years, Doctoral dissertation, University of Melbourne.

Appendix

Angela's picture narration task response

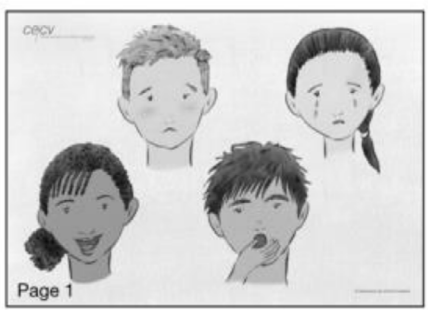
No attempt has been made to rectify grammatical errors in the French transcriptions, and these have been carried through into the English translations in order to give a sense of the error rate in the student's speech.

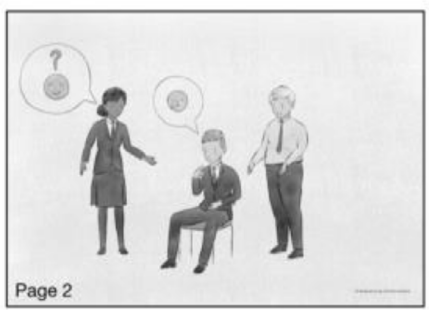
Term 4, 2019: Number of Word Types = 74, Average Sentence Length = 6, Number of Key Points = 10, Number of Words per Minute = 44

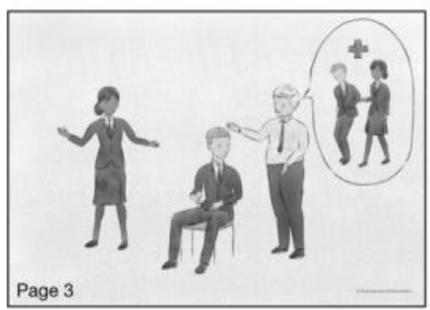
Response: Aujourd'hui, il très mal. Il veut un infermerie. Aujourd'hui, il très mal. Il veut va infermerie. Elle triste. Elle triste, triste. Elle pleut. Elle très triste, elle pleut. Bonjour, ça va? Je suis très content mais j'ai fatigué aussi, tu? Bonjour merci mais j'ai très malade. Je veux vomir dans le poubelle. Est-ce que tu peux va le professeur s'il vous plaît? Bonjour, est-ce que tu veux? Bonjour le professeur. Il trés malade. Est-ce que je aller à l'infirmerie avec il? Oui, merci, va vite le infermerie. Bonjour il, Bonjour tout le monde, comment ça va? Ça va bien merci, et toi? Mais j'ai aussi, suis fatigué et mal. Bonjour professeur. Est-ce que tu peux va le infermerie avec elle? Elle très mal et très fatiguée; merci. Comment ça va? Je suis bien J'ai très bien aussi, merci. mais il très fâché. Je ne sais pas. Je pense que il m'a … est fâché parce que il très fatigué. Oui, je suis aussi fatigué.

[Today, he very hurt. He wants a sick bay. Today, he very hurt. He wants go sick bay. She sad. She sad, sad. She is crying. She very sad, she is crying. Hello, how are you? I am very happy but I have tired also, you? Hello, thank you but I have very sick. I want to vomit in the rubbish bin. Can you go the teacher please? Hello, you want? Hello the teacher. He very sick. Do I to go to sick bay with he? Yes, thank you, go quickly the sick bay. Hello he, hello everyone, how are you? I'm well thank you, and you? But I have also, am tired and hurt. Hello teacher. Can you go the sickbay with her? She very hurt and very tired, thank you. How are you? I am good. I have very good also, thank you. But he very angry. I don't know. I think that he … is angry because he very tired. Yes, I am also tired.]

Prompt images:







7 Designing and implementing an assessment course for English language teachers: Insights into assessment literacy development

Frank Giraldo
Departamento de Lenguas Extranjeras, Universidad de Caldas, Colombia

This chapter describes the development and implementation of a course designed to foster the assessment literacy of a group of 18 in-service English language teachers from state high schools in Colombia. It highlights:

- A multidimensional approach for characterizing LAL before course design
- The nature of the course
- Its impact on teachers' LAL development
- A call for research and resources on ethics and fairness for teachers' LAL

Este capítulo describe el desarrollo e implementación de un curso diseñado para fomentar la competencia en evaluación de idiomas (*Language Assessment Literacy*) de un grupo de 18 docentes de inglés en ejercicio, provenientes de escuelas secundarias estatales en Colombia. El capítulo resalta los siguientes puntos:

- Un enfoque multidimensional para caracterizar la LEL antes de diseñar el curso
- Características del curso
- Su impacto en el desarrollo de la LEL de los docentes
- Un llamado a la investigación y recursos sobre ética y justicia para la LEL de los docentes

Introduction and context

For the past 20 years in the field of language testing, language assessment literacy (LAL) has received scholarly attention. There has been a major emphasis on defining what LAL entails and how it relates to different stakeholders involved in language assessment (Brindley 2001, Davies 2008, Fulcher 2012, Kremmel and Harding 2020, Taylor 2013). Among these stakeholders, teachers remain a key group whose practices and needs in language assessment have been studied extensively (Fulcher 2012, Sayyadi 2022, Sultana 2019, Vogt and Tsagari 2014).

Regarding teachers' professional development in LAL, Giraldo (2021) reports that assessment courses for pre- and in-service teachers have an overall positive impact on their LAL. Specifically, teachers become aware of practice and theory that are fundamental for sound assessment. Additionally, as Giraldo (2021) explains, research on these types of courses focuses on the outcomes of training, rather than how LAL was developed. This is an under-researched area in LAL: although this construct has been clarified, to some extent, approaches to developing it – i.e., pedagogies of LAL – require further research and discussion (Fulcher 2020). Against this background, the purpose of the present case study is to report on a pedagogical experience involving the planning and implementation of an assessment course for language teachers.

Thirty in-service English language teachers from state high schools (both urban and rural) in the central region of Colombia willingly participated in a research project that had a training component. Both the research project and the training component, henceforth referred to as the LAL course, are described in this chapter. In general, English teachers in this country are expected to plan, teach and assess the English language based on a national language learning policy called *La Guía 22* [Guide 22] (Ministerio de Educación Nacional de Colombia 2006). This policy describes standards for learning English in the skills of listening, reading, writing, spoken monologue and conversation; the standards for all grades (1st to 11th) are adapted from the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001).

After the launch of the standards, the Ministry of Education published a document with guidelines for language teaching and suggested syllabuses for all grades in elementary and high schools; this set of documents is commonly called The Suggested Curriculum (Ministerio de Educación Nacional de Colombia 2016). Thus, English teachers in Colombia are expected to use these standards and the suggested curriculum in their schools. For assessment, there is a general document that establishes principles and procedures for assessment in all subjects, called Decree 1290 (Ministerio de Educación Nacional de Colombia 2009). However, there is no document

to orient *language* assessment; the suggested curriculum briefly defines competence-based assessment and assessment *for* and *of* learning (Ministerio de Educación Nacional de Colombia 2016). In general, the teachers in the LAL course reported in this chapter are free as to how they want to assess their students (i.e., with the assessment instruments they deem appropriate), based on the standards. Thus, in the English class in general, teachers assess language performance in the four skills.

In Colombia, there is a national test that assesses various subject areas, including English. The test, called Saber 11, assesses reading and grammar and vocabulary in context. There is some evidence that this test leads to washback among English language teachers in Colombia, but this research is underreported (Barletta and May 2006).

The LAL course was part of a research project called *Literacidad en Evaluación de Lenguas Extranjeras y Desarrollo Profesional* [Language Assessment Literacy and Professional Development], sponsored by the Research Office (Vicerrectoría de Investigaciones y Posgrados) at Universidad de Caldas in Manizales, Colombia. The project was divided into two major stages: a diagnostic, to plan the course, and an implementation, i.e., the LAL course itself.

To design the course, all 30 teachers participated in the diagnosis; however, 18 participated in the course; the remaining 12 could not do so because of mental and physical health issues associated with COVID-19. In the diagnostic, the teachers completed a questionnaire to identify their learning needs in assessment; the instrument included 32 closed-ended items with topics about language assessment and one open item for teachers to suggest other topics for the course. The teachers' feedback was used for course planning and implementation. A key question in the questionnaire was whether they had taken any language assessment courses or modules: 20 had never studied language assessment, eight had taken brief modules, and two had taken complete courses during their initial pre-service education. In fact, in Colombia, pre-service teachers' LAL is still in its infancy; studies have reported a lack of assessment competence development during pre-service teacher education (Giraldo and Murcia 2019, Herrera and Macías 2015).

The LAL course

In 2019, I wrote a research proposal to design and implement the language assessment course reported in this chapter. This proposal went through a research office (similar to an Institutional Review Board) and was evaluated by two professors in my country. Once the proposal was accepted, between 2020 and 2021, I designed and taught the assessment course, which was free for the participating teachers. The course covered theoretical, practical, and critical issues in language assessment. The teachers and I, the course

instructor, met twice a week for two hours over 10 weeks. Table 1 includes major characteristics of the course.

Instructional sequence

Before the two workshops each week, the teachers read LAL information in a handbook that I wrote for the course. The Appendix includes excerpts from this 10-module handbook; each module consisted of definitions and issues in the topic for the week, practical examples connected to language assessment in the Colombian context, and one or two tasks for the teachers to bring to the week's workshops.

During synchronous sessions, the teachers shared their task answers and participated in test analysis, test planning, and discussions. During the second workshop of the week, the teachers had a major task, which was the design of a language assessment by writing test specifications and deriving an instrument from them. Lastly, the teachers completed a short questionnaire on the perceptions they had towards the course and, overall, how it was impacting their professional development.

Table 1 Characteristics of the LAL course

Characteristic	Description
Delivery	Online through Google Meet
Length	10 weeks Two workshops every week; two hours each 40 hours total
LAL topics	Purposes and qualities for language assessment Assessing speaking, writing, listening, reading, and integrated skills Alternative assessment Ethics and fairness in language assessment Online assessment
LAL materials	A handbook as preparatory reading
LAL tasks	Discussions around test use and consequences Designing instruments for language skills, often in teams Brainstorming ideas for planning assessments Analysing assessment instruments and their overall usefulness

Research design, data collection and analysis

The research objective for the diagnostic was to describe the teachers' LAL by characterizing their learning needs and their skills in instrument design. For data collection and analysis, the research team[1] used a mixed-methods

1 The research team consisted of four research assistants who were undergraduate students of a BA in Modern Languages and me as the principal investigator. The assistants' role was to collect, organize, and analyse data for the project.

approach. The teachers completed a questionnaire where they selected from a variety of LAL topics to be included in the course. These data were used for content selection in the diagnostic stage (see Table 3). Also, they participated in an individual online interview and shared two assessment instruments, one for reading or listening and one for speaking or writing. We analysed the assessment instruments against design guidelines to see whether they presented any design difficulties. All the data from the diagnostic stage were scrutinised to determine the content and some instructional strategies for the LAL course.

The research objective for the implementation was to determine teachers' perceptions of course components and its impact on professional development. The data came from a three-point questionnaire with Likert-type items and a focus group interview in which teachers shared their insights into the course and impact on their professional development.

For all the qualitative data in this study, the research team used a thematic coding approach (Creswell 2016, Kuckartz 2014) to look for general trends that could be useful for planning the course (i.e., diagnostic) and for evaluating its impact on teachers (i.e., implementation). In the research team, we discussed differences in coding until we reached consensus involving major trends. Table 2 summarizes the research approach we used.

Table 2 Research methodology

Project stage	Data collection	Purposes	Analysis
Diagnostic	• Questionnaire on LAL learning needs • Individual interview • Analysis of assessment instruments	Needs analysis: Characterize LAL and plan the course	• Percentages to rank topics • Trends in teachers' answers and problematic areas regarding instrument design
Implementation	• Questionnaire on content and activities; professional development • Focus group interview	Evaluation: Determine teachers' perceptions towards the course and its impact on their LAL development	• Percentages to identify perceptions of content and activities • Themes regarding impact on professional development

Findings

This case study reports the findings in two sections. The first section presents those from the diagnostic stage, and the second presents the findings from the implementation stage of the course.

Table 3 Learning needs in language assessment

Topics	Yes	No
Reliability	97% (29)	3% (1)
Practicality	97% (29)	3% (1)
Assessment in virtual environments	97% (29)	3% (1)
Designing assessments for speaking and writing	94% (28)	6% (2)
Assessing speaking	94% (28)	6% (2)
Assessing writing	94% (28)	6% (2)
Assessing integrated skills	94% (28)	6% (2)
Assessing pronunciation	90% (27)	10% (3)
Validity	90% (27)	10% (3)
Authenticity	90% (27)	10% (3)
Impact	90% (27)	10% (3)
Assessing listening	90% (27)	10% (3)
Communicative language testing	90% (27)	10% (3)
Formative/Alternative assessment	87% (26)	13% (4)
Designing assessment specifications	87% (26)	13% (4)
Designing and using self-assessment	87% (26)	13% (4)
Assessing reading	84% (25)	16% (5)
Task-based assessment	84% (25)	16% (5)
Assessing vocabulary	84% (25)	16% (5)
Doing bilingual assessment	84% (25)	16% (5)
Giving feedback	81% (24)	19% (6)
Assessing cultural aspects	81% (24)	19% (6)
Designing and using peer assessment	81% (24)	19% (6)
Validation	81% (24)	19% (6)
Evaluating existing assessment instruments	77% (23)	23% (7)
Using and interpreting statistics	74% (22)	26% (8)
Assessing grammar	74% (22)	26% (8)
Designing and using portfolios	74% (22)	26% (8)
Fairness	71% (21)	29% (9)
Designing test items (multiple choice, true and false, matching)	65% (19)	35% (11)
History of language testing and assessment	52% (16)	48% (14)
Ethics	7% (2)	93% (28)

Diagnostic stage

Determining LAL needs from various angles

In the diagnostic stage, 30 teachers completed a questionnaire, written in English, on language assessment content they felt they needed to learn about in the course. Their answers in this instrument allowed us, the research team, to prioritize topics that would become part of the handbook and the workshops in the course. Table 3 includes a ranking of the topics the teachers chose as needs for LAL development.

The results in the questionnaire suggest that teachers wanted to learn about virtually everything that was included in the instrument. This is an issue that Fulcher (2012) warns about, arguing that teachers may find all topics about language assessment important, especially when the respondents willingly want to be part of an LAL project. The teachers also participated

in an interview before they completed the questionnaire; to a certain extent, the interview corroborated the results in the questionnaire. A trend in the answers was that they wanted to learn about as much as the course could offer. The sample below, translated from Spanish, illustrates this trend.

> T29
> Me gustaría aprender de todo lo que me puedas enseñar … cómo construir evaluaciones que sean pertinentes, que apunten a lo que se supone deben apuntar. Me gustaría aprender sobre todo.
> [I would like to learn about everything you can teach me … how to build assessments that are pertinent, that aim at what they are supposed to aim. I would like to learn about everything.]

The data from the questionnaire and the interview confirmed what Fulcher (2012) explains. However, perhaps the teachers could not clearly, or thoroughly, articulate what about assessment they really needed (I discuss this further below). Thus, to gain a more complete picture of teachers' LAL, we asked them to share two assessment instruments to enable a more direct observation of teachers' strengths and weaknesses. As we analysed these documents, we found two major issues: threats to construct validity and limited authenticity in design. For example, the test item below was an instrument designed to check 8th grade students' reading comprehension. However, the item can be answered without reading, as common knowledge will suffice. This issue was evident throughout the listening and reading tests the teachers shared.

> Ins23 ReadingMCQ2
> 2. What are the primary colors to make green?
> a.- blue
> b.- blue and yellow
> c.- yellow
> d.- green

Also, it should be noted that options *a* and *c* are partially correct and that *d* can be eliminated based on logic. In conclusion, the analysis of assessment instruments led us to determine that the teachers needed to learn about guidelines for item design to improve the usefulness of tests to assess, in this case, reading or listening skills, not other construct-irrelevant factors.

Authenticity was something that seemed problematic in the assessments. In one of the speaking tests, students were asked to listen to their teacher and repeat what he says in a recording. This assessment task, arguably, is not assessing language the way it is used in real life. Additionally, the rubric

included construct-irrelevant components, i.e., visual aids, which may lead to variance not attributed to learners' speaking skills, hence another threat to construct validity.

Implementation stage

The overarching importance of construct definition

During the focus group interview, the teachers remarked on how important it is to them to clearly define the construct for an assessment. For example:

> T12 – FGI
> Primero que todo, es fundamental tener un constructo … es la piedra angular y sobre esa piedra, se planean las actividades, las metodologías o las estrategias, para que así el estudiante pueda desarrollar esa competencia.
>
> [First, having a construct is basic … it's the cornerstone and on that stone, activities, methodologies or strategies are planned, so that the student can develop that competence.]

In the questionnaire results, 15 teachers reported that during the week they studied about key questions in language assessment (the Why, the What, and the How), there was considerable learning on constructs. Fourteen teachers, out of fifteen, stated that they learned a lot about this topic, while the remaining teacher stated they learned a little. In fact, construct definition, as a key assessment component, emerged in teachers' answers regarding practice in language assessment, as it will become apparent in the next finding.

Rigor and care in language assessment

The teachers also reported that, during the course, they became more aware that designing an assessment requires accurate planning and attention to detail; it is not a haphazard process, with the constructs considered as central elements in design. The sample below comes from the focus group interview:

> T20 – FGI
> La evaluación no es simplemente poner preguntas al azar sino evaluar habilidades … y no solo basarnos en las 4 habilidades, sino que es evaluar de manera precisa y concisa una sub-habilidad, que es el constructo … ver un proceso tan detallado y meticuloso, nos damos cuenta de que dejamos cosas de lado que son importantes. Vemos el paso a paso del qué y cómo evaluar.
>
> [Assessment is not just to put questions randomly but assess skills … and not just basing ourselves on the four skills, but assess accurately and

concisely a sub-skill, which is the construct … seeing such a meticulous and detailed process, we realize we leave aside things that are important. We see the step by step of the what and how to assess.]

Discussion

Because most teachers stated that they had never studied language assessment, the results in the diagnostic stage seem expected. However, asking teachers what they want to learn about assessment may lead them to state they want to learn about everything being offered; this issue happened both in the questionnaire and the interview. Thus, I consider this to be a limitation in this study. Particularly in the questionnaire, teachers could choose what they needed to learn, but this instrument used technical terminology. This begs the question on how respondents can determine whether they need to learn something they may or may not know about. This issue may be exacerbated when teachers, in an interview, do not have the terminology to articulate their thoughts and state what they want to learn about assessment. Asking stakeholders what areas of LAL they want to explore may become a LAL test in itself. Thus, I believe we have a challenge for LAL research: how can we ask stakeholders what they want to learn about language assessment, without using technical terminology?

On the other hand, the third method in the diagnostic stage (instrument analysis) provided concrete data on areas for improvement in the teachers' LAL. Through analysing their assessment instruments, it was possible to determine technical areas to be addressed in the course, namely item writing and rubric design, along with the qualities of validity and authenticity.

The teachers' stated lack of LAL training, and corresponding needs for learning, are also findings in Fulcher (2012), Sayyadi (2022) and Vogt and Tsagari (2014). The fact that two teachers wanted to be in the course, despite having learned about language assessment in other scenarios, shows support for the argument that LAL is about development rather than an end state, a current topic in LAL debates (Baker 2021, Yan and Fan 2020).

Overall, the data gathered from the diagnostic stage, especially the instrument analysis, provided useful feedback for the kind of LAL course that would most benefit the teachers. They expected a course addressing knowledge and skills in LAL, which is partially in line with what authors suggest in their LAL descriptions for language teachers (Fulcher 2012, Inbar-Lourie 2008, 2012, Taylor 2013).

As for the findings in the implementation stage, the teachers were continually engaged in writing up assessment specifications for designing tests. Test specifications need careful consideration, and construct definition is a key element in them. Thus, specifications served as a bridge to connect theoretical ideas about the What in language assessment; careful, rigorous planning; and

well-crafted instruments. This practice-based approach, which Fulcher (2020) calls a task-based pedagogy for LAL, highlights the importance of instilling LAL courses with practical activities for learning about language assessment (Fulcher 2012, Giraldo and Murcia 2019, Giraldo 2021). Such activities may seem more useful when the local assessment context is considered: the LAL course in this initiative was designed for a particular group of teachers with specific learning needs and assessment lifeworlds.

Finally, I must address a concern that emerged in the course. In my reflections as the course instructor, I realized that I could not address the principles component of LAL, i.e., ethics and specifically fairness, given the lack of pedagogical materials for these topics in *classroom* language assessment. Thus, during a discussion in a workshop, I asked the teachers what unfair practices they had seen or conducted. Below is what two teachers shared in this discussion:

> T7
> He visto algunos profesores, al final del periodo, semestre o lo que sea. Nosotros los profes debemos reportar … si los estudiantes están perdiendo o no. Podemos ver como algunos profesores que ni siquiera evalúan … ¡no tienen notas! [La profesora se refiere a un compañero] "¿El estudiando está perdiendo otras clases? OK, entonces la mía también."
>
> [I've seen some teachers, at the end of the term, semester, whatever. We teachers must report … if students are failing or not. We can see some teachers who don't even assess … they don't have grades! [T7 talks about a colleague by quoting them] "Is the student failing other subjects? OK, mine, too."]
>
> T13
> Uno de mis estudiantes se quebró su pierna porque estaba corriendo en la escuela … los coordinadores y el rector les pidieron a todos los profesores que él tenía que pasar todas las materias, porque si no pasaba, los papas iban a denunciar a la escuela.
>
> [One of my students broke his leg because he was running in the school … the coordinators and principal told all teachers that he had to pass all the subjects, because if he didn't pass, the student's parents would sue the school.]

The anecdotal comments above have something in common: construct irrelevance – the teachers had to assign grades that did not reflect language constructs. In general education, this is called score pollution (Rasooli, Zandi and DeLuca 2019). Ethics and fairness are discussed in high-stakes assessment, not in *classroom* language assessment. Thus, LAL discussions might need to consider two research avenues: how to design materials which

address principles in LAL – especially in classroom language assessment – and how to teach this topic in assessment courses for language teachers.

To summarize, two major findings regarding the teachers' professional development in LAL emerged during implementation. On the one hand, the teachers emphasized the fundamental role that constructs play in the assessment process; they also reported that designing high-quality assessments requires care and rigor. Although this was a case study with research features and findings intrinsically relevant, as a teacher educator, I believe LAL courses should prioritize conscientious design, fuelled by careful specification writing, and guided by adherence to guidelines for test construction. My experience led me to believe that a practice- or task-based approach to LAL development is largely beneficial for teachers' LAL.

References

Baker, B (2021) *Unpacking the concept of language assessment literacy for all key stakeholders*, paper presented at New Directions 2021, Bogotá, Colombia, 1–3 July 2021.

Barletta, N and May, O (2006) Washback of the ICFES Exam: A Case Study of Two Schools in the Departamento del Atlántico, *Íkala: Revista de Lenguaje y Cultura* 11 (1), 235–261.

Brindley, G (2001) Language assessment and professional development, 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, 126–136.

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

Creswell, J W (2016) *30 Essential Skills for the Qualitative Researcher*, Thousand Oaks: Sage Publications.

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.

Fulcher, G (2020) Operationalizing Assessment Literacy, in Tsagari, D (Ed) *Language Assessment Literacy: From Theory to Practice*, Newcastle: Cambridge Scholars Publishing, 8–28.

Giraldo, F (2021) Language assessment literacy and teachers' professional development: A review of the literature, *Profile: Issues in Teachers' Professional Development* 23 (2), 265–279.

Giraldo, F and Murcia, D (2019) Language assessment literacy and the professional development of pre-service language teachers, *Colombian Applied Linguistics Journal* 21 (2), 243–259.

Herrera, L and Macías, D (2015) A call for language assessment literacy in the education and development of teachers of English as a foreign language, *Colombian Applied Linguistics Journal* 17 (2), 302–312.

Inbar-Lourie, O (2008) Constructing a language assessment knowledge base: A focus on language assessment courses, *Language Testing* 25 (3), 385–402.
Inbar-Lourie, O (2012) Language assessment literacy, in Chapelle, C A (Ed) *The Encyclopedia of Applied Linguistics*, Hoboken: Wiley-Blackwell, 2,923–2,931.
Kremmel, B and Harding, L (2020) Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the language assessment literacy survey, *Language Assessment Quarterly* 17 (1), 100–120.
Kuckartz, U (2014) *Qualitative Text Analysis: A Guide to Methods, Practice and Using Software*, Thousand Oaks: Sage Publications.
Ministerio de Educación Nacional de Colombia (2006) *Estándares Básicos de Competencias en Lenguas Extranjeras: Inglés. Formar en Lenguas Extranjeras: ¡el reto! Lo que Necesitamos Saber y Saber Hacer*, Bogotá: Imprenta Nacional.
Ministerio de Educación Nacional de Colombia (2009) *Fundamentaciones y orientaciones para la implementación del Decreto 1290 de 2009*, Bogotá: Ministerio de Educación Nacional de Colombia.
Ministerio de Educación Nacional de Colombia (2016) *Pedagogical Principles and Guidelines: Suggested English Curriculum*, Bogotá: Team Toon Studio.
Rasooli, A, Zandi, H and DeLuca, C (2019) Conceptualising fairness in classroom assessment: exploring the value of organisational justice theory, *Assessment in Education: Principles, Policy & Practice* 26 (5), 584–611.
Sayyadi, A (2022) In-service university-level EFL instructors' language assessment literacy and training needs, *Profile: Issues in Teachers' Professional Development* 24 (1), 77–95.
Sultana, N (2019) Language assessment literacy: An uncharted area for the English language teachers in Bangladesh, *Language Testing in Asia* 9 (1), 2–14.
Taylor, L (2013) Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections, *Language Testing* 30 (3), 403–412.
Vogt, K and Tsagari, D (2014) Assessment literacy of foreign language teachers: Findings of a European study, *Language Assessment Quarterly* 11 (4), 374–402.
Yan, X and Fan, J (2020) "Am I qualified to be a language tester?": Understanding the development of language assessment literacy across three stakeholder groups, *Language Testing* 38 (2), 219–246.

Appendix

Excerpts with sample content from the handbook the teachers used

This is some of the theory the teachers read:

The What
This is, perhaps, the most important question that has driven language testing and assessment for many (many!) years. The What refers to *constructs* for language assessment.

What is a construct?
A construct is the specific skill (or set of skills) about which we want to collect information through assessment. Constructs are not visible, so we need to use instruments to stimulate them so we can 'see' them. For example, listening comprehension is a construct we cannot see directly, so we use an assessment instrument to know what a student's listening comprehension may be like.

Language ability
In language assessment, the 'mama construct' is called 'language ability'. You have probably seen this concept also called 'communicative competence'; this is the construct that underlies *La Guía* 22 [Guide 22] and the Suggested Curriculum. We, humans, 'show' or 'utilize' our communicative competence as we listen, read, speak, or write in any language. This is why developing language ability – or communicative competence – is the quintessential goal of language teaching, learning, and assessment.

This is a task the teachers had to do before a workshop. The task is aligned with the general language learning policy in Colombia:

> What methods do you use to collect information about your students' skills in the English language? Are they useful to collect information about how your students are developing the standards for a specific grade? Make a list of methods you use and bring it to Week 1.

In this excerpt, teachers could see examples of communicative tasks for reading.

Texts	Purposes for reading these texts
Phone notifications	(De)Activate something; update something; evaluate whether the text is useful or just an ad; reply to someone; read and store information for an appointment.
Phone reminders	Do a task.
Bills	Find the fee one has to pay; identify details regarding our consumption.
Receipts	Know how much to pay or whether a tip needs to be included; check whether the price meets the items one consumed.
Plane tickets (before and after printing)	Confirm whether all information is accurate; as the time for the flight approaches, read it to confirm time and gate.
Everyday texts (menus, ads, etc.)	Get information; solve specific needs, as in a doctor's prescription; go around places, as in street signs; behave accordingly, as in traffic signs.
Newspaper ads or articles	Read to get information and be up to date; for entertainment; to find useful information, e.g., a car on sale.
Specialized articles	Learn about a particular subject: What it is, what its characteristics are, what is going on with a topic; and many more details.
Literature: poems, novels, etc.	For pleasure. In academic contexts, people read these texts to learn about something, to write reports on them, and to analyze their aesthetics.

8 Putting LAL into practice: What happens in classrooms and why?

Susan Sheehan
The University of Huddersfield, UK

Vivien Berry
(formerly) British Council, London, UK

Based on the belief that the classroom is an under-represented venue for research into teachers' language assessment literacy (LAL), this chapter presents a qualitative case study of six non-subject-specific English for Academic Purposes (EAP) teachers in a language institute associated with a British university. It includes:

- A description of the setting for the case study including participating teachers
- The development and trialling of a schedule for observation of assessment activities
- Data analysis of the frequency of different assessment activities
- Teachers' explanations of their rationale for using these assessment activities
- A discussion of the usefulness of the observation schedule and suggestions for further use

Introduction and background

Teachers have often been surveyed by researchers who seek to understand language assessment literacy (LAL) by asking what teachers know about assessment and/or what they would like to know more about. Whilst this research has provided interesting insights into aspects of language teachers' assessment literacy, in the words of Jiang in her 2020 study of teachers' questioning in the classroom: 'studies directly observing teacher AL in their classroom practice are sparse' (2020:3).

However, empirical research into classroom-based assessment practices can inform teacher educator practice, and facilitate the development and understanding of how LAL is conceptualised and operationalised by

teachers of English as a foreign language (EFL). It can also complement and triangulate with self-report measures from other studies. Consequently, it can be argued that the classroom is under-represented in LAL research and more of this type of work is necessary. In order to compensate to some extent for this under-representation, the focus of the study described in this chapter is specifically related to LAL in the classroom. It was inspired by Colby-Kelly and Turner's (2007) study of formative assessment practices, which shows how data can be collected from such practices regarding what is happening in the classroom and the likely thinking behind those actions.

Our qualitative investigation into how LAL is manifested in classroom practices was conducted with a small group of teachers working within a language institute. We were particularly interested in discovering: 1) what classroom-based assessment activities the teachers used; 2) what their motivation was for using these activities; 3) if there was a relationship between institutional assessments and teachers' classroom practices.

Observational and post-interview methodology – a commonly used data collection method in qualitative research designs (Curdt-Christiansen 2019) – was selected as the most appropriate approach, and is described below.

Methods

Setting

The study was conducted at an English Language Centre associated with a university in the north of England. The English courses provided by this centre prepare students for degree courses at the associated university and others in the UK, and include aspects of both English for Academic Purposes (EAP) and General English. The observed classes covered non-subject-specific EAP. Assessments included a range of activities which were similar to task types commonly used at UK universities, such as giving presentations and compiling annotated bibliographies, and occurred at fixed intervals over a year. As the institution had rolling enrolment, teachers prepared students who had differing levels of experience of the assessment tasks.

Participants

The study used a sample of convenience consisting of six British teachers from the language centre who volunteered to take part in the project. All six were highly experienced and had worked in the UK, while three had worked overseas. They were also relatively highly qualified, with three holding MA degrees from UK universities. Table 1 summarises the

Table 1 Teachers' qualifications and experience

Teacher	TEFL qualifications	TEFL work experience	TEFL years of practice
1. (Alice)	CELTA DELTA Master's in TESOL	EFL teaching in Turkey EAP teaching in UK universities	22
2. (Peter)	CELTA DELTA Master's in TESOL	EFL teaching in South Korea EAP teaching in UK universities	8
3. (Grace)	Trinity Certificate Trinity Diploma	ESOL* teaching in the UK EAP teaching in UK universities	9
4. (Emily)	DELTA Master's in Education	ESOL teaching in Australia EAP teaching in UK universities	8
5. (Chloe)	CELTA DELTA	ESOL in UK schools EAP teaching in UK universities	35
6. (Hazel)	CELTA DELTA	ESOL in UK schools EAP teaching in UK universities	43

**English for Speakers of Other Languages.*

participant teachers' qualifications and teaching experience. All names are pseudonyms.

Materials and data collection

For our study, the researchers developed an observation schedule following Colby-Kelly and Turner's (2007) study of assessment for learning (AFL) practices. Our observation schedule contained 16 different assessment activities as shown in Table 2.

The schedule also included space where we could record any additional information relating to the observed assessment activities, which were divided into three-minute sections. The time at the start of each section was noted in the time section on the schedule. We placed a tick (✓) next to the activities observed. If the same activity was observed for a second time, a second tick would be added, and so on. The notes section was used to record comments made by the teachers or note student reactions to the assessment activities. A researcher adopted the role of 'non-participant observer' (Curdt-Christiansen 2019:338), seated unobtrusively out of the eyeline of the students; we felt this best facilitated natural teacher–student interactions. Every three minutes a new observation schedule was started. The total time of each observation was 60 minutes, therefore there are 20 completed schedules per observation.

The observations were conducted by the research team member who worked closest to the language institute. Distance made it impractical for

Table 2 Observation schedule of assessment activities

	Time		Notes
	Activity	Tick	
1.	Communicating assessment results to students		
2.	Questioning during talk – impromptu discussions, presentations, group work, etc.		
3.	Spontaneous feedback during talk – positive/negative, evaluation, assistance, scaffolding etc.		
4.	Co-construction of meaning/topic/form		
5.	Organise and manage assessment process		
6.	Bureaucratic – to fulfil obligations to an external agency		
7.	Pedagogic – to inform teacher and support professionals on learner progress		
8.	Learning – to support, as opposed to measuring learning		
9.	Focus on form		
10.	Learning aims and objectives		
11.	Subject contents and activities that need to be given emphasis		
12.	Judgment on whether or not learning aims have been achieved		
13.	Task-based assessment		
14.	Alternative assessments – portfolios, reflective journals, self- and peer assessment, scaffolding students' learning		
15.	Reporting and giving feedback in the assessment process		
16.	Using strategies to enhance learning		

both researchers to observe the lessons or participate in the post-observation interviews. The observation schedule was piloted with colleagues of the research team and was revised a number of times. The observer and the colleagues who piloted the schedule were highly experienced in conducting observations within their roles as teacher educators. This wealth of experience informed the development of the schedule. When the schedule was used, it was found to be broad enough to cover all the assessment activities observed, and the space for notes allowed the observer to capture a lot of classroom activity in a short amount of time. Some of the observation sheets were blank as not all sections featured assessments.

As soon as teaching schedules allowed, we conducted stimulated reflection interviews with the six teachers to understand and/or clarify the motivation for their observed classroom assessment practices. In these interviews, we invited the teachers to discuss the observed assessment activities and the thinking behind them. The interview prompts included examples of the language they had used in the classroom, such as the questions the teachers had used in the lesson. Other prompts were descriptions of teacher activity, such as when they had been observed to manoeuvre around the classroom.

Results

Data analysis

The data analysis process was conducted in two stages. In the first stage, the observation schedules were collated to establish which assessment activities were most used across all the observations (see Table 3). For each of the activities shown in Table 3, we give examples of teachers' classroom talk in italics.

Table 3 Most frequently observed assessment activities

	Activity	Rate
1.	Spontaneous feedback during talk – impromptu discussions, presentations, group work, etc. *'Yes, good, But you're missing a word there. Don't worry about referencing now.'*	93
2.	Co-construction of meaning/topic/form *'If you're unsure look for keywords. As a clue look for place names, negative phrases.'*	51
3.	Subject contents and activities that need to be given emphasis '*You have to learn to pick bits that support your argument. The aim is to put the points in your own words.'*	50
4.	Alternative assessments – portfolios, reflective journals, self- and peer assessment, scaffolding students' learning *'Give feedback on presentations, both positive and negative and you have to give at least one positive and give a point to improve on. Make sure it's constructive criticism.'*	42
5.	Using strategies to enhance learning '*Reading aloud can help you get the grammar. It's a good thing to do. Just remember in the exam you won't have time to write all that – you need to pick key points.'*	39
6.	Organise and manage assessment process *'In the exam, where do you take your ideas from? You have to take ideas from the text and paraphrase it.'*	33
7.	Focus on form *'With **enhance** you have to have a noun – you can't have that here – it's always followed by a noun.'*	33
8.	Reporting and giving feedback on the assessment process *'Last term you weren't clear in your essays. This activity has shown you how important topic sentences are.'*	30
9.	Learning – to support, as opposed to measuring learning *'This is not an essay, so you don't have to do that. Can you see how these sentences give the general idea?'*	26
10.	Questioning during talk – impromptu discussions, presentations, group work, etc. *'You're just copying from the text – you need to change it. What does that mean? So, active games can what?*	26

The interview data were then thematically analysed with a focus on the comments made by the teachers which explained the thinking behind the observed teaching activities. The following three key themes were identified: movement, autonomy and institutional assessments. The first relates to assessment moments happening when the teachers were circulating. The second concerns the teachers' use of assessment to promote learner autonomy. The third covers the relationship between classroom-based assessments and the institutions' assessments.

For the top three activities, we also provide examples from the observations and comments from the interviews in which the teachers provide a rationale for the activity.

Activity 1: Spontaneous feedback during talk – impromptu discussions, presentations, group work, etc.

Emily made several comments whilst she was walking around the students who were in groups at tables arranged in islands. For example, she commented on one student's work: 'Yes, exactly, yes, write that, write what you said. Pick out the points you want – you can cut things out.' When asked about the thinking behind the comment she explained: 'Because I think everything that we've done, we've been taught has an element of assessment, because you're always assessing students, as soon as they walk in the door, not formally maybe, but informally.'

Peter was observed giving spontaneous feedback in a similar way to Emily as he manoeuvred around the classroom. He stated that at one level such moving around the classroom ensured the students stayed on task. Beyond this a deeper reason for this monitoring was to identify students who were struggling or to find common areas of difficulty in the group. When working with one group of students he made the following comment: 'You need to pay extra attention, Elisa, you were asking about this.' He was aware that this student had experienced difficulties with this language point in a previous session and so was careful to ensure that the student in question felt encouraged to engage with the lesson content. In the interview Peter clarified: 'So, if I approach both [strong and weak] students and ensure they are on task by asking some questions regarding what we're doing, and they are able to answer them, it shows that they've understood instructions and that they're making progress through tasks.' Another feature of Peter's assessment practice was making a note on the whiteboard of problematic language so he could work through it with the whole group later in the session: '… if it's speaking, for example … and when monitoring that I may note information, I put it on the board and save it until the end of the class as deferred feedback.'

Echoing Peter's remarks about using monitoring, Grace commented: 'but also I just find that when you walk round, they know you are looking they will work more, some of them work hard but some of them won't unless

you are stood over them, but I also like to see how they are getting on.' The spontaneous feedback she made included praise such as: 'Well done!' Another comment indicated to the student that they needed to change a part of their work: 'You are missing a word there though.'

Hazel's spontaneous feedback comments were mainly positive. She made frequent use of praise. For example, she told one student: 'That's beautiful work.' To another she said: 'Lovely. You're doing really well.' This use of praise created a very positive atmosphere in the classroom and the students responded well to it. In the interview Hazel highlighted the importance of monitoring and individual feedback in ensuring that each student received attention and relevant feedback. She explained: 'We have to put them into groups, but I'm interested in finding out really what each person can do.'

Chloe's spontaneous feedback comments were also generally positive: 'Yes, that's right, rather than asking am I right? Think first and then you'll have more confidence in yourself.' She described the work of another student as being: 'Brilliant.' She described such monitoring and feedback as a cornerstone of her assessment practice, commenting that: 'I have to continually assess, whilst I'm in the classroom, checking around the tables and making sure what they're doing.' Chloe's comments to students were not limited to praise. We also observed her correcting pronunciation of the word *quarter* and asking questions to make the students aware when their language use was problematic such as: 'Decline, is decline up? No! It isn't.'

Alice was sparing in her use of praise and her comments were more of a challenge to the students to improve their performance. She told one group of students: 'Ask yourselves – what do you need to move forward? Done OK overall – but this is wrong you can do better.' In the interview Alice gave the following explanation about these remarks: 'So, I wanted the students to question themselves, so I will have been asking questions for them to ask themselves, questions I want them to ask themselves but that might not automatically pop into their heads.'

Activity 2: Co-construction of meaning/topic/form

Whilst co-construction is generally taken to mean that the teacher and students were building understanding together, in this study we use the term to show that the students were not explicitly told things but were brought to an understanding through inductive teaching and practices such as dialogue and questioning. As with the assessment activity of spontaneous feedback, the activity of co-constructing meaning, topic or form was mainly observed when the teachers were walking around the classroom and interacting with small groups or individual students.

For Chloe, one of the reasons for using this activity was to help the students to consider themselves as successful language learners. In the observation she made the following comment: 'I'd have that in a

different place.' This was the start of a dialogue between Chloe and a student about the best place to insert a sentence in a paragraph and the reasons why the student's initial choice of place for the sentence was not the best. In order to avoid students feeling unduly negative after an assessment activity, she focused on guiding the students to be better language users without worrying about less successful aspects of their performance. In the interview Chloe stated: 'And it's just to try and show them how to get it better, rather than what they've done wrong.'

Grace also used the co-construction activity whilst monitoring the students. We observed her asking multiple questions which related to the paragraphing activity the students were completing. For example, she pressed a student with the following comment: 'You're right. I want to know why it is right.' She nudged students who had given a wrong answer towards a better answer as she wanted them to be able to write in paragraphs in an upcoming written assessment. She provided the following rationale for this assessment activity: '… to make sure they have understood the lesson material and that they can apply it to what they are going to have to do.'

Hazel used questions to guide students to develop their understanding of form. She asked students questions such as: 'What's wrong with this one? What should it have?' and 'Can you tell me why this is wrong?' We did not observe Hazel explicitly correcting the students. She co-constructed with them an understanding of form.

Chloe, Grace, and Hazel worked with students to co-construct meaning and form whereas Peter co-constructed the topic. Throughout his lesson on presentations, Peter created a dialogue with the students about the key features of a presentation. His elicitation and questioning techniques guided the students through the stages of presenting and delivering a presentation. Peter framed these discussions as a *route-map* and used it so the students would arrive at the destination of delivering a good assessment when required to do so as part of the end-of-course assessments. He shared the presentation marking criteria with the students, which they used when engaged in peer assessment. Peter explained his decision to share the criteria thus: 'So, I try to demystify the end goal for them and share the criteria, so they understand where we are now and what we have to achieve in the future then this helps.'

Similarly, Alice used dialogue to co-construct an understanding of how to improve their performance in the institutional assessments. At the start of the lesson, she returned to the students' marked examinations and initiated a discussion with the class on the strengths and weaknesses of the papers. On the board she drew a model of reflection which outlined a cycle of reflection and action planning and stated that it was imperative that the students used the papers as a prompt for reflection. For Alice, reflection is fundamental to language learning. In her interview Alice lamented that students tended to fixate on the score and this fixation hindered a more profound

engagement with the process of reflection and acting following reflection. She commented: '… you need to reflect on performance … it's not just the score.' Like Peter, Alice wanted the students to take on more responsibility for their learning. In Alice's case she wanted this sense of responsibility to extend to other areas of their lives beyond English language learning. She clarified: '… because if you are going to be successful as a learner and in life you have to reflect.'

Activity 3: Subject contents and activities that need to be given emphasis

This assessment activity can be viewed as a bridge between the classroom-based assessment activities and those of the institution. The teachers made clear links between classroom activity and the institutional assessments. The teachers did not create these assessments but were involved in their administration and marking. They used the institutional assessments as a jumping off point for reflection or for other types of assessment such as peer-assessment.

Chloe made several comments when teaching the whole group which related to the institutional assessments. For example, she told them: 'This will help you with your similes for writing' and '… you will need to paraphrase – you will need to do this in your essay.' In the interview she stressed that such comments were made with the intention of helping the students to understand the requirements of the assessment so as to perform better in them.

Grace related the paragraphing activity which accounted for most of the observed lesson to the writing assessments which the students were due to complete. She stated: 'with the paragraphing they are currently in the process of writing a main body paragraph for the final essay that they'll do … so any assessment that I do in class needs to check that they've got the basic idea of that structure.' She also suggested that being able to identify topic sentences, which was part of the paragraphing activity, would help the students to perform better in assessments of reading.

Of all the teachers we observed, Peter's lesson contained the most examples of this assessment activity. His lesson was focused on presentation skills as the students would have to deliver a presentation as part of their end-of-course assessments. He gave a deliberately poor presentation and asked the students to identify its weak points. He then gave the students guidelines about the stages of presentation preparation. The students used these guidelines as they wrote notes for their presentations. Peter shared the presentation assessment criteria with the class. Working in pairs the students used the criteria to provide feedback on the presentation notes. He provided the following rationale for the sequence of activities outlined above: 'I guess I was trying to encourage more critical thinking as well. And yes, trying to engage them in peer feedback, try and create

noticing and give them the responsibility as opposed to taking all the responsibility myself.' So, the activities were designed to promote critical thinking and to enable the students to assume more responsibility in the learning process.

Emily's lesson also contained numerous examples of links being made between classroom activity and the end-of-course assessments. For example, she elicited from the group the ways in which an argument essay differed from the other types of essay which they had studied. A further example of Emily relating class activity to examination activity is this comment: 'Just remember in the exam you won't have time to write all that – you need to pick key points.' In the interview Emily stated that giving such emphasis on the end-of-course assessments would not normally be a feature of her teaching and assessment practices. Emily also suggested that she felt the need to talk about the assessments in detail as she found the group to be very laid back in their approach to assessments. Assessment practices are influenced by the personality of the group and the relationship between the teacher and the students. She summarised her view of the group thus: 'I kind of have that sort of confidence now that they know what to do, they understand what to do, and then I believe that they'll do it. So, with that extra push, but possibly not as stressful a push as it would be if they were panicking.'

Discussion and conclusions

From the data discussed above we determined the following features which seemed to be present in the assessment practices of all the teachers we observed. Firstly, assessment often occurred when the teachers were moving around the classroom monitoring the students. These assessments were spontaneous and were focused on guiding the students through the learning process. This spontaneous feedback was used to monitor progress and provide further support to weaker students. This would seem to suggest that classroom-based assessment practices are student-centred and responsive to the needs of individuals or at most, small groups of students.

Secondly, the assessment practices were used as a tool to promote reflection and learner autonomy. The teachers wanted the students to take a more active role in the language learning process and to ensure that they were capable of reflecting on their own performance and so reduce their reliance on teacher judgment. They also wanted the students to develop learning strategies which would be of use to them both in the English language classroom and beyond.

Thirdly, the teachers used the institutional assessments as a springboard for their classroom-based assessment practices. They did not use mock examinations or similar test preparation activities. Rather, they used activities such as peer assessment to help the learners to assimilate the marking criteria

used by the institution and so develop a deeper understanding of what was required of them and what they needed to do to meet these requirements. Thus, an activity which started as one which focused on institutional assessment requirements, progressed into one which focused on peer assessment. Such a transformation suggests that classroom assessments are multi-functional and can address different student needs.

In conclusion, the evidence from these classroom observations leads us to believe that the language assessment literacy of the six teachers we observed was highly developed at a practical level as they all successfully deployed a range of assessment activities and could eloquently articulate their rationale for using them. Each teacher was able to use the information generated by the assessment activities to create an individualised learning experience for each student, which we consider to be laudable in the context of a course designed to be taught in groups.

The observation schedule itself was a useful tool which allowed us to capture a range of assessment activities that were relevant to the context. It helped the observer to maintain focus on assessment activities and so avoid being drawn into other aspects of the lesson such as board work or classroom management. The extensive piloting allowed the observer to feel confident when using the schedule.

However, the schedule would need to be adapted for use in different contexts. For example, an observation of a class of young learners would require it to be revised to include more age-appropriate assessment activities. If other researchers were to use the schedule, they would also need training to develop an appropriate level of confidence. In addition, teams of observers would need training to ensure the assessment activities were being documented consistently across the team.

References

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.

Curdt-Christiansen, X L (2019) Observations and field notes: recording lived experiences, in McKinley, J and Rose, H (Eds) *The Routledge Handbook of Research Methods in Applied Linguistics*, New York: Routledge, 336–347.

Jiang, Y (2020) Teacher classroom questioning practice and assessment literacy: case studies of four English language teachers in Chinese universities, *Frontiers in Education* 5, available online: doi: 10.3389/feduc.2020.00023

9 Providing continuing professional development through a language assessment literacy MOOC

Carolyn Westbrook, British Council, UK

Richard Spiby, British Council, UK

In recognition of the need for assessment literacy among language teachers, the British Council's *Language Assessment in the Classroom* Massive Open Online Course (MOOC) was created with the aim of promoting language assessment literacy primarily among teachers but also other stakeholders. This case study outlines:

- The contents of the MOOC
- Participants' background, perceived understanding of language assessment concepts and satisfaction with the course
- Insights gained from running the course over a four-year period
- Lessons learned and recommendations for others wishing to develop similar MOOCs

Introduction

Assessment should play a crucial role in teacher education (Popham 2009) but, in reality, teachers often carry out assessment duties without adequate training (Stiggins 2014). Consequently, many researchers have called for language assessment literacy (LAL) training to become an integral part of teacher training (Herrera and Macías 2015, Stiggins 2014). Nowadays, more and more teacher training courses include assessment literacy, yet many teachers still have to rely on Continuous Professional Development (CPD) opportunities to develop their assessment literacy.

Literature review

Giraldo (2021) found that LAL training can take various forms. He presents two types of typical initiatives to improve language teachers'

assessment literacy: self-access materials and professional development programmes. Self-access materials include textbooks on language testing, journals and other online resources, whereas professional development programmes include blended learning courses, workshops (of a few hours in duration), short courses (e.g. 1–3 weeks), longer courses (e.g. one semester), ongoing training (lasting for several years), and action research projects.

A different option would be to offer training through a MOOC. There are several types of these but a common one is the XMOOC (usually shortened to MOOC), which includes content and videos for participants to engage with independently, but instruction primarily directed by the course educator(s) (Lowenthal, Snelson and Perkins 2018). MOOCs have several advantages over traditional training options: they are online, often low-cost or free, and are flexible so participants can study when and where they wish. Furthermore, in contrast to free online resources, such as the British Council's 'How Language Assessment Works' (British Council 2022) or Trinity College London's 'Assessment Literacy' resources (Trinity College London 2022), they allow for interaction with educators.

Apart from the advantages outlined above, there are a number of motivational factors leading to participation in MOOCs. Hakami, White and Chakaveh's (2017) review of the literature on MOOCs found that three types of motivational factors contributed to participation: learner-related, institution and educator-related, and platform and course-related factors. The first of these includes personal aspects including the learner's attitude and intrinsic/extrinsic motivation; social aspects such as interaction with learners; and factors related to educational or professional development, including the relevance to their job, their opportunity to improve their skills and access to opportunities which would otherwise not be available to them. The institution and educator-related factors were linked to the perceived reputation of the institution and the interaction with the educators. Finally, platform and course-related factors include the perceived usefulness and ease of use of the platform, the course content, and the popularity of the MOOC.

Despite the many advantages of MOOCs, retention on these courses is often an issue. Alraimi, Zo and Ciganek (2015) cite several studies which suggest that the average completion rate for MOOCs is less than 10% of enrolments, while Reich (2014) found that, out of nine HarvardX courses, only 22% of the participants who intended to complete the course actually did so. Conversely, Hone and El Said (2016) found that 32.2% (n=122) of their participants completed the MOOC they registered for, which appears to be a relatively high completion rate. These studies highlight several factors which contribute to retention: the perceived reputation of the provider

and perceived openness[1] of the course by participants (Alraimi et al 2015), perceived quality of course content and opportunity for interaction with the educators (Hone and El Said 2016), and learner intention to complete (Reich 2014).

The Language Assessment in the Classroom MOOC

In recognition of the need for low-cost, flexible and widely available assessment literacy training, the *Language Assessment in the Classroom* MOOC was created to promote LAL primarily among teachers but also other stakeholders, such as teacher trainers and administrators. The MOOC is a free, four-week course designed to increase participants' understanding of the principles of language assessment and to equip them with knowledge and practical skills (Fulcher 2012) for developing effective assessment instruments in a classroom context. Considering the challenges to retention outlined above, the course was designed to be highly engaging through a variety of input materials, tasks and resources, while frequent educator interaction with participants was also prioritised. To the best of the authors' knowledge, this is the only MOOC focusing on LAL.

Course topic coverage was decided on the basis of practicality from a teacher's perspective. Weeks 1–3 are based on assessment of the four skills and language knowledge. Although this may be seen as a traditional approach, we felt that this would give the course a recognisable structure and be accessible to participants from a wide range of educational backgrounds. Week 4 covers more specialised areas of test development and evaluation with the focus on implementation in a classroom context. Inevitably, many topics had to be excluded. For example, theoretical topics and wider issues of ethics and test impact were considered out of scope in such a classroom-focused course, while other topics, such as fairness and accommodations, could be covered as part of wider educational training. The main topics covered in the course are:

- Assessing speaking and writing
- Assessing reading and listening
- Assessing vocabulary and grammar
- Assessment and course planning
- Test design
- Assessing integrated skills

1 Openness relates to 'freedom of access to educational content' (Paulsen 1993 in Alraimi et al 2015).

- Assessing young learners
- Assessing online
- Developing and evaluating tests
- Item analysis

Each week of the course comprised a number of steps. Steps are small units of the course focusing on a specific piece of content and based on a single type of input. The number of steps changed slightly as amendments were made to the course but in the fifth run, Week 1 comprised 24 steps, Week 2 included 21 steps, and Weeks 3 and 4 each had 17 steps. Step input was delivered through texts, videos, recorded interviews with language testing experts, and animations. Quizzes, discussion tasks and practical tasks, such as scoring tests and calculating basic statistics, enabled participants to engage with the input. The educators moderated the discussions and responded to participants' comments throughout the course. The lead educators also hosted Facebook Live sessions, in which participants posed questions live to renowned language testing experts, and all educators discussed key themes arising from learners' comments during the week in end-of-week videos.

The course ran five times between April 2018 and November 2021. Table 1 below shows the participation data for the five runs. Over that period, there were 31,199 learners from over 170 countries worldwide. As there are different levels of participation on MOOCs, the table shows the data for the number of people who signed up for the course (joiners) compared to those who actually started (learners), the number of learners who actively participated in the steps (active learners), those who posted comments (social learners), and the number of learners who completed at least 50% of the

Table 1 MOOC participation data (April 2018–November 2021)

Run no	Joiners	No. of countries	Learners	Active learners	Social learners	Learners with >50% step completion	Learners with >90% step completion
1	6,645	151	4,350	2,525	1,076	666	428
2	10,348	158	5,805	3,107	1,216	901	602
3	21,294	170	12,409	7,142	3,473	3,221	2,459
4	8,836	165	5,408	2,907	1,329	954	668
5	5,755	150	3,227	1,643	837	606	465
Total	**52,878**		**31,199**	**17,324**	**7,931**	**6,348**	**4,622**
Mean	**10,575.6**	**161**	**6,239.8**	**3,464.8**	**1,586.2**	**1,269.6**	**924.4**
%			**59.0%**	**55.5%**	**25.4%**	**20.3%**	**14.8%**

The mean for each column is the total number of participants divided by the number of runs (n = 5). The percentage for the 'learners' column is calculated by dividing the total number of learners by the total number of joiners. The percentages for the other columns are the total numbers as a percentage of the total number of learners.

steps and at least 90% respectively. As seen in Table 1, 59% (n = 31,199) of joiners (n = 52,878) started the course and the number of learners per run ranged from 3,227 on Run 5 to 12,409 for Run 3, which took place in April–May 2020 (at the peak of the global pandemic). The number of participants completing at least 90% of the steps ranged from 9.8% (n = 428) of learners on the first run to 19.8% (n = 2459) on Run 3 with an average retention rate across all five runs of 14.8%.

Table 2 below shows the total numbers of comments posted by all learners and educators on the discussion tasks by week and in total on each run of the MOOC. The mean is the total number of comments divided by the number of steps in that week. In total, 126,470 comments were posted by participants and moderators over the five runs of the MOOC, with the third run accounting for almost half of these. This run was unusual as this course took place at the height of the global pandemic when people were unable to access in-person professional development due to lockdowns.

Table 2 Means and totals for comments posted during the five runs

Run no	Week 1		Week 2		Week 3		Week 4		Overall total
	Total	Mean	Total	Mean	Total	Mean	Total	Mean	
1	6,463	269.29	3,261	125.42	2,014	100.70	1,676	83.80	13,414
2	8,109	368.59	4,258	202.76	2,643	155.47	2,123	132.69	17,133
3	26,051	1,184.14	16,479	784.71	11,164	620.22	8,681	542.56	62,375
4	9,773	444.23	5,576	265.52	3,426	190.33	2,662	166.38	21,437
5	5,500	250.00	3,136	149.33	1,863	109.59	1,612	94.82	12,111
Grand total	**55,896**		**32,710**		**21,110**		**16,754**		**126,470**

Catering for participants in a global context

While a great advantage of a MOOC is its reach, this also poses the challenge of defining the context in which it is used. Global participation across over 170 countries means that it is impossible to have knowledge of all participants' backgrounds, making it difficult to anticipate the different needs of such a diverse range of learners. As seen in the questionnaire results below, participants are involved in a range of roles in the education sector. The implications of this for the MOOC in terms of accessibility are clear. Firstly, a threshold of language ability for participants cannot be guaranteed. The language level of the course content was largely kept to B1 of the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001), since we considered this the lowest level at which assessment principles could be clearly articulated, yet still maintain accessibility. Secondly, since a minimum level of training for participants cannot be assumed, technical terms, theoretical concepts and

detailed abstract explanations were avoided unless strictly necessary to understand fundamental concepts in assessment, such as those listed in Figure 4.

Despite such challenges, the variety of contexts is a source of interest and a rich area for discussion. As context of use is a critical factor in developing and implementing assessment instruments, and integral to the validity of assessment (Chalhoub-Deville and O'Sullivan 2020), such discussions were an important element of participant comments. Teachers commented not only on issues in their own context but also attempted to frame the same issue in terms of colleagues working in very different conditions. Consequently, several participants adopted a problem-solving approach to find the most appropriate test for the context. More generally, participants could see how their counterparts solve assessment problems under different constraints. The educator's role during this process is to ensure that discussions are constructive and conducted according to sound assessment principles.

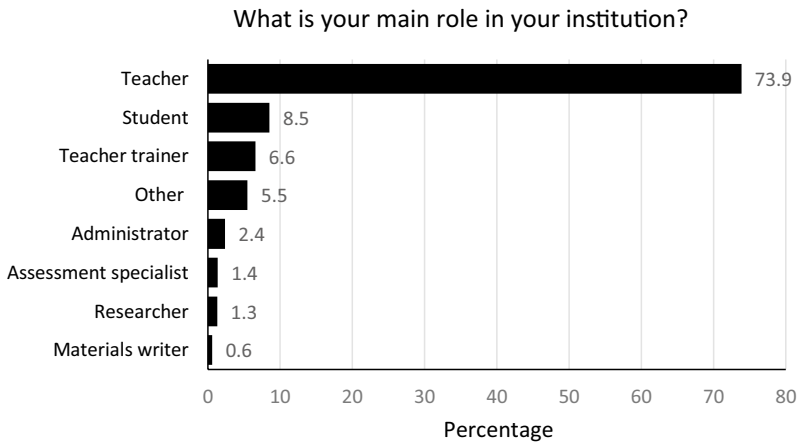
Evidence suggests that such discussions helped create a virtual community of practice within this diverse population. Participant comments indicated that communication between peers did take place with participants learning from each other. However, there was the risk of erroneous information being passed between participants, necessitating intervention from educators to make corrections, either directly through replies to comments or in end-of-week videos, where misunderstandings could be dealt with in greater detail. This was resource-intensive, but clarification of certain important concepts, such as teaching to the construct, was largely successful in halting the spread of confusion and provided opportunities for learning.

Questionnaire results

Pre- and post-course questionnaires were administered to course participants. The purpose of the pre-course questionnaire was to discover the profile of participants, in terms of knowledge and experience of language assessment, as well as motivations and expectations. The aim of the post-course questionnaire was to evaluate different aspects of the course and determine to what extent the MOOC had addressed participant needs. Findings from both questionnaires informed the subsequent approach of educators and influenced course content. The survey results reported here are taken from the three most recent iterations of the course. Data were collected through a voluntary online survey tool using links provided to participants in course steps. A total of 7,784 responded to the pre-course survey, while 2,003 responded to the post-course survey. Selected findings, focusing on participants' professional background, experience of assessment and satisfaction with the course, are presented below.

In terms of respondents' backgrounds, the largest proportion (36.7%) were involved in secondary education, followed by higher education (26.0%), with a smaller number in the primary sector (15.2%). Over a fifth of participants (22.1%) were at other non-specified institutions. The vast majority of respondents (Figure 1) were teachers, with a small number of students and teacher trainers. Overall, these findings show that the participants represented the target demographic of the course.

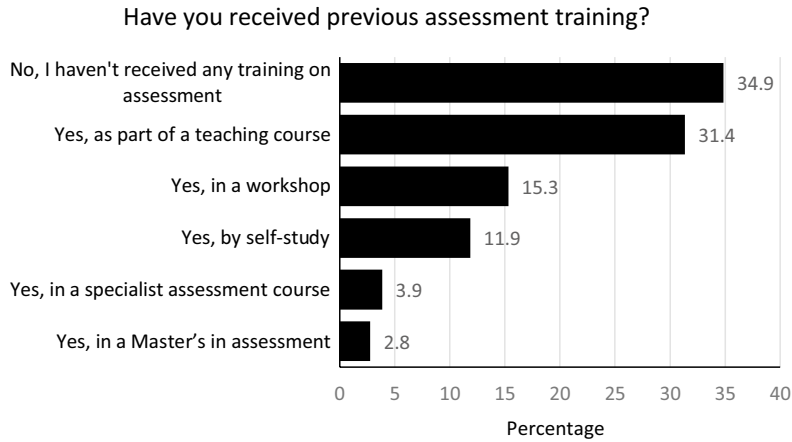
Figure 1 Pre-course employment: Role



About a third of respondents (34.9%) had not received any assessment training before joining the MOOC, while slightly fewer (31.4%) had had training as part of their teaching course (Figure 2). Workshops and self-study accounted for most of the rest. This large proportion of participants without previous assessment training would appear to support Stiggins' concerns about teachers (2014) mentioned in the Introduction, but of course they are precisely the intended audience for the course.

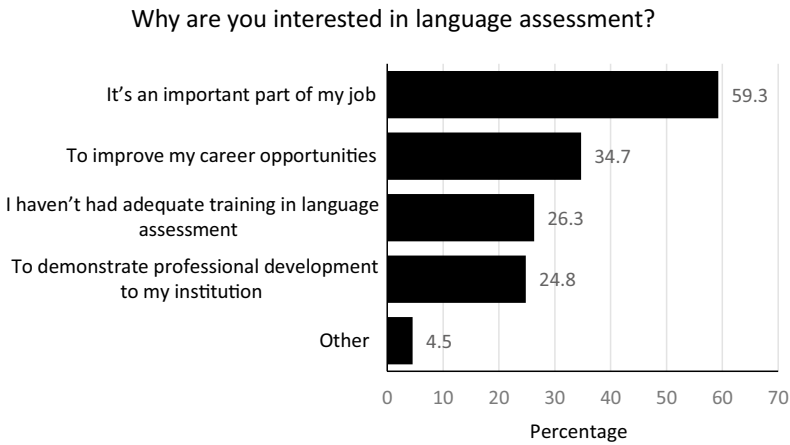
Despite a relative lack of training, the majority of respondents (69.9%) reported that they had been actively engaged in developing tests. Thus, it can be assumed that significant numbers of teachers and other educationalists are expected to produce assessment materials without the appropriate background training, indicating a real need for large-scale assessment literacy projects. This high level of involvement with test development is also reflected in the reasons for respondents' interest in the subject (Figure 3). As several authors have noted (Stiggins 2014, Stoynoff and Coombe 2012), assessment can take up a considerable amount of a teacher's time, and this importance was reported by most respondents (59.3%), implying a need to better inform current professional practice, as well as improving career

Figure 2 Pre-course assessment training



prospects (34.7%). About a quarter (26.3%) believed that they had not had sufficient training in their career so far.

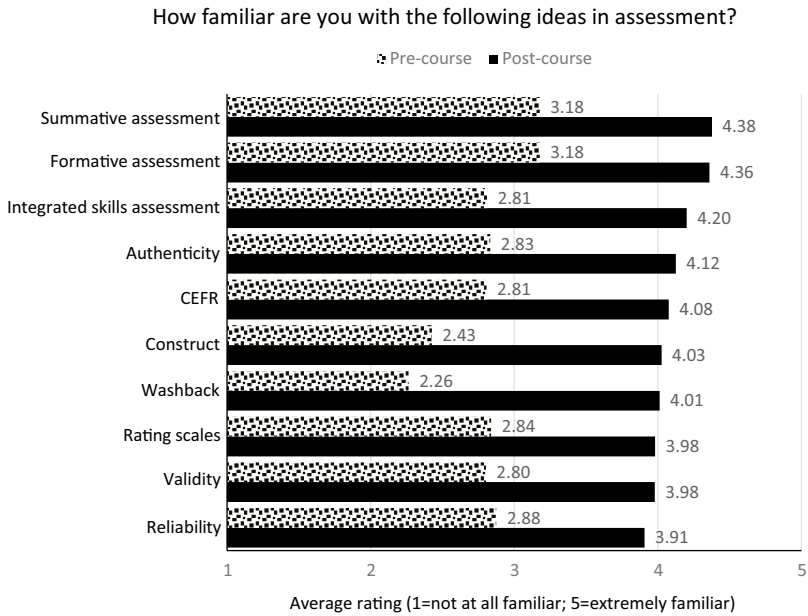
Figure 3 Pre-course interest in language assessment



As noted above, effort was made to reduce the number of technical terms used, as we believe that the excessive use of specialist technical language can be frustrating. Nonetheless, several fundamental terms were used, with potentially unfamiliar items shown in Figure 4. While there appeared to be some familiarity with concepts such as formative (3.18) and summative

(3.18) assessment, participants appeared largely unfamiliar with others. Construct (2.43) and washback (2.26), both of considerable relevance to teaching practice, seemed particularly less well known. At the end of the course, respondents reported much higher levels of familiarity with the same concepts, again with relatively high ratings for formative (4.36) and summative assessment (4.38), and large gains for construct (4.03) and washback (4.01). However, lower levels of familiarity were given for validity and reliability, which may reflect in part the complexity of these ideas and also the sparing use of these terms during the course.

Figure 4 Pre- and post-course assessment concept familiarity



At the end of the course, participants were asked about their overall experience during the MOOC. Figure 5 shows reasons for the use of the MOOC as a professional development opportunity. In line with Hone and El Said (2016), the most popular reason (82.4%) was the flexibility of the learning mode. This was closely followed by interest in the topic (76.9%), which is considered an important motivation to finishing the course (Hakami et al 2017, Hone and El Said 2016, Milligan and Littlejohn 2017). Smaller but notable numbers of respondents cited lack of access to other forms of CPD (22.9%) and self-access materials (16%), supporting the findings of Odden et al (2002, in Misra 2018).

Figure 5 Post-course reasons for MOOC participation

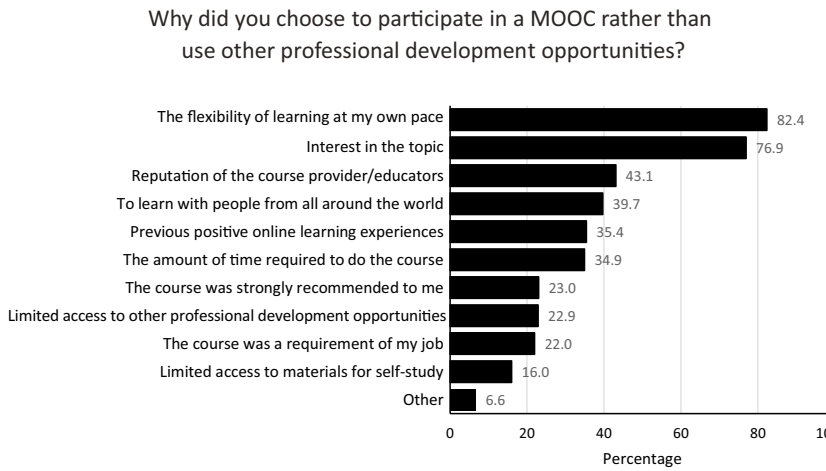
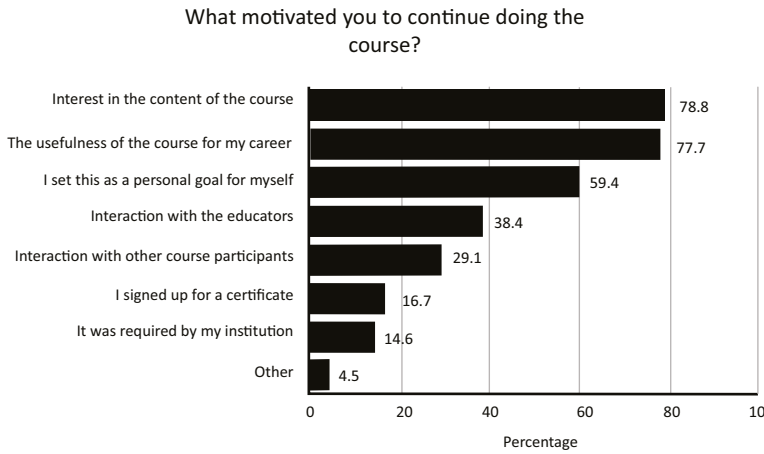


Figure 6 Post-course motivational factors



When participants were specifically asked about motivation (Figure 6), intrinsic interest in the content area was again a primary factor (78.8%), alongside the more instrumental factor of career benefit (77.7%). Interaction with the educators played a greater role (38.4%) than interaction with other course participants. This reflects the fact that educators made a concerted effort to engage through comments and end-of-week videos since instructor interaction has been found to have a positive impact on course completion (Hakami et al 2017, Hone and El Said 2016). However, other participants

were less of a motivating factor, supporting the impression that interaction among participants is challenging for MOOC course creators.

The *Language Assessment in the Classroom* MOOC comprises different resources and learners were asked to report on their usefulness (Figure 7). As reported previously (Fulcher 2012, Giraldo 2020) with such items, the responses display little variation, making interpretation difficult. Most features were rated above 4.3 out of 5, with links to external websites, including that of the British Council and other test providers, and instructional videos being considered particularly useful. End-of-week videos were perceived as less useful. Comments about these suggested that the greater length of these videos and Facebook Live broadcasts (3.84) created issues for internet bandwidth, so quick and

Figure 7 Post-course learning resources

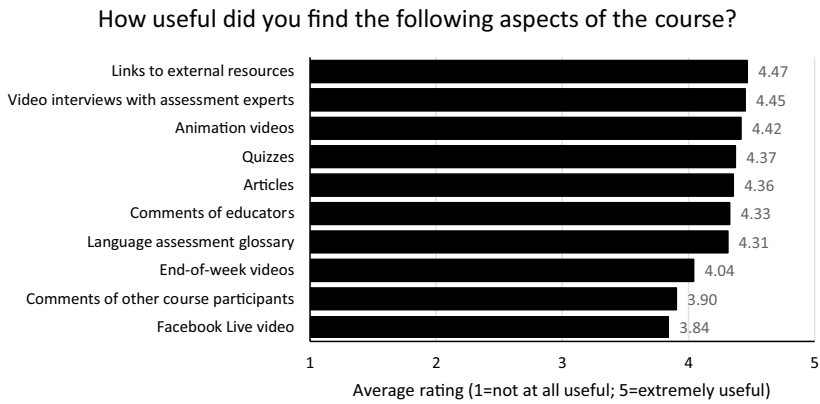
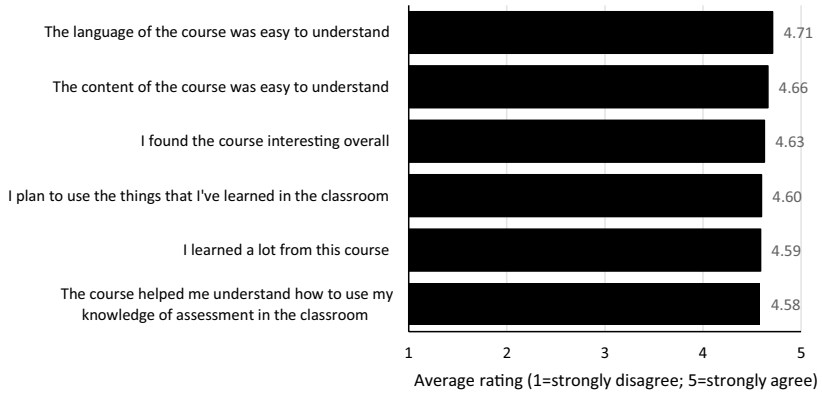


Figure 8 Post-course learning and application



easy to access resources were favoured by course participants. Consistent with Figure 6, educator comments were considered more useful than those of course participants although the latter was still rated quite highly.

One of the key indicators of the success of any course is the level of participant engagement. As such, feedback was encouraging (Figure 8). Respondents reported high levels of interest (4.63) and learning (4.59). One concern in the development of the course had been to find a balance between accessibility and informativeness with regard to language and content. Results suggest that this was largely successful. Additionally, respondents had a high level of agreement with statements related to the use and implementation of course content in professional life. While this is no guarantee of actual impact in the classroom, it does show that participants felt equipped to apply new ideas to their own contexts.

Discussion

In this section, we will outline some lessons learned from both the literature and our findings from running this MOOC and offer advice for others wanting to create MOOCs for teacher professional development.

Cost

Cost has been found to be a prohibiting factor (Odden et al 2002, in Misra 2018), so our MOOC was free of charge for the duration of the four-week course, which might have contributed to participation. However, running a MOOC with a high level of educator interaction proved to be costly. Fortunately, the platform we used offered the option for participants to pay for an upgrade, which allows ongoing access to the course materials after the course has finished. Consequently, the British Council was able to recover some of the costs from the upgrades purchased. We would therefore recommend that, if possible, MOOCs should be offered free of charge or for very low cost for an initial period to widen opportunities for participation. However, institutions should consider platforms that enable participants to purchase upgrades as these may bring in third-stream income without the threat to accessibility, which occurs if there is a course fee from the outset.

Time commitment

Running the MOOC was not without its challenges. While gauging the amount of time participants needed to work through the input was relatively easy, the amount of time required to read all the comments varied considerably. Therefore, managing the time commitment from educators

(and participants) was quite difficult initially. Monitoring the interaction was important because, if the course had not been actively moderated, participants' questions/misunderstandings would not have been addressed, which, in turn, may have decreased their satisfaction level and impacted negatively on their learning and, possibly, retention. After the first run, however, it was possible to see how much interaction there had been in each step. This information allowed the lead educators to divide up educators' time approximately evenly by allocating a range of steps to each educator. In this way, educators were able to follow the interaction only on the steps they were responsible for, thereby reducing their time commitment and the running costs of the MOOC. We would therefore suggest dividing up the steps between the educators to reduce costs and lower the time commitment but we would stress that, initially, at least, it is necessary to monitor how much interaction each step is eliciting, and to be ready to reallocate educators to steps which are eliciting more interaction than others.

Knowledge base

Since the vast majority of participants on the course were non-native speakers of English, setting the language level and complexity of the content at the appropriate level were important considerations. We opted for CEFR Level B1 for the input as this meant that participants would have a reasonable command of the language, which was sufficient for us to be able to explain difficult concepts in a fairly simplistic manner but without losing the complexity of the topic. This appeared to be a good level as the questionnaire results suggest that participants found both the language and the content easy to understand, which may have also contributed to retention. Comprehension was likely supported by the glossary of terms provided at the beginning of the course along with an initial task aimed at familiarising participants with the glossary and key terms. Many participants commented that they found the document very helpful and learned about new concepts that they had previously not been aware of thanks to the initial familiarisation task. Moreover, from our point of view, this ensured that all participants had a minimum level of understanding at the outset.

In addition, giving participants a pre-course questionnaire to find out about their backgrounds and level of knowledge may have contributed to retention as participants became aware of areas they were not yet familiar with. This may have provided motivation for the participants while also enabling educators to be aware of where issues may arise.

Motivation and retention

Flexibility and interest in the topic were the two most widely cited reasons for participating in the MOOC while the value of the course for participants' careers was also frequently cited. These factors are in line with the research into motivational factors for studying on MOOCs (Hakami et al 2017, Milligan and Littlejohn 2017) and may have aided retention.

Similarly, participants rated the interaction with the educators highly. Therefore, as suggested by Hakami et al (2017), having educators on hand throughout the course run may also have increased participants' motivation and contributed to retention. Although this interaction was relatively expensive, it appeared to help promote interaction and the exchange of ideas between participants and educators. Participants appeared to very much appreciate this as indicated by their 'likes' on educator comments. In addition, educators were able to step into the conversation when misunderstandings arose.

Another aspect of motivation and retention from the literature was related to the period of time during which participants dropped out of MOOCs. Hone and El Said (2016) found that most dropouts happened before or at the mid-point. With this in mind, we decided to have a 'hook' to keep people interested in Week 3, that is, after the mid-point of the course. This was a topic which we thought would be of particular interest to participants. For example, in the last three courses since the pandemic, online assessment was selected as a topic for one of the live sessions, which we pushed as a highlight of the course. Although we were not able to collect detailed weekly dropout figures for the course, it appears that, based on the decrease in the numbers of comments posted across the four weeks of each run of the course, most dropouts appear to have happened within the first week with fewer dropouts in Week 2 and fewer again in Weeks 3 and 4 respectively.

We recommend that the scheduling of live sessions or content which is considered a highlight of the course is planned carefully. Having a live session or a key content area soon after the mid-point may be a way to maintain motivation and aid retention. Additionally, for live broadcasts, it should be ensured that participants from as many different time zones as possible can participate. If possible, we would strongly recommend having access to back-end staff who can deal with technical issues during the course and live sessions, and, if necessary, help with the technical aspects of course design.

The reputation of the provider and educators

As shown in Figure 5 above, the reputation of the provider was the third most popular reason for choosing to study on the MOOC. This suggests that the British Council is fortunate to enjoy a good reputation worldwide already.

However, for less well-known institutions, running a successful MOOC may increase their visibility and serve as a marketing tool (Lowenthal et al 2018) which, in turn, can raise the profile of the institution and enhance its reputation. For the educators, MOOCs also provide an opportunity to not only pass on their knowledge via the content of the MOOC but also to spread the word about the research areas they are working on, thus making their research more visible (Blackmon 2016, Lowenthal et al 2018). Therefore, for anyone interested in offering a MOOC, it is worth raising awareness of the benefits for the providing institution and the educators within your own institution to gain buy-in from senior management and attract funding for your MOOC.

Professional development for both participants and educators

Clearly, MOOCs serve a professional development purpose for participants. However, educators benefit from their involvement too. When participants share insights about their contexts and concerns, this serves as a professional development opportunity for both educators and other participants as they all learn about different contexts around the world. We would therefore urge educators to see MOOCs as a professional development opportunity for themselves as well as the participants and to encourage participants to share information about their contexts for the benefit of all involved.

Reach and impact

As the name suggests, MOOCs are intended to have large numbers of participants. While this meant that the reach of our MOOC was substantial, with over 31,000 participants from over 170 countries, the size of the course inevitably meant that a lot of participants remained anonymous. Consequently, it was difficult to know who exactly the course was reaching and which institutions were benefitting from it. Similarly, it was virtually impossible to measure the impact of the MOOC on classroom practice. Therefore, to measure the impact of the course, we would suggest requesting permission to collect participants' contact details in order to contact them after the course to carry out a follow-up impact study.

Conclusion

The British Council's *Language Assessment in the Classroom* MOOC was a very successful course, receiving very positive feedback from participants. This was due in no small part to the very competent course development team and

highly experienced back-end staff, who guided us on the technical side, as well as the knowledgeable and enthusiastic educators who helped us run the course.

There were challenges and we cannot deny that a good deal of work was involved in running the MOOC. However, we would advocate the use of MOOCs for teachers' continuous professional development and we hope that, by giving access to free, flexible, professional development, we have contributed to improving assessment literacy among teachers around the world.

References

Alraimi, K M, Zo, H and Ciganek, A P (2015) Understanding the MOOCs continuance: the role of openness and reputation, *Computers and Education* 80, 28–38.

Blackmon, S J (2016) Through the MOOCing glass: professors' perspectives on the future of MOOCs in higher education, *New Directions for Future Research* 167, 87–101.

British Council (2022) *How language assessment works*, available online: www.britishcouncil.org/exam/aptis/research/assessment-literacy

Chalhoub-Deville, M and O'Sullivan, B (2020) *Validity: Theoretical Development and Integrated Arguments*, British Council Monograph Series 3, Sheffield: Equinox Publishing.

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

Fulcher, G (2012) Assessment literacy for the language classroom, *Language Assessment Quarterly* 9 (2), 113–132.

Giraldo, F (2020) A post-positivist and interpretive approach to researching teachers' language assessment literacy, *Profile: Issues in Teachers' Professional Development* 22 (1), 189–200.

Giraldo, F (2021) A reflection on initiatives for teachers' professional development through language assessment literacy, *Profile: Issues in Teachers' Professional Development* 23 (1), 197–213.

Hakami, N, White, S and Chakaveh, S (2017) Motivational factors that influence the use of a MOOC: learners' perspectives, *Proceedings of the 9th International Conference on Computer Supported Education (CSEDU 2017)* 2, 323–331.

Herrera, L and Macías, D (2015) A call for language assessment literacy in the education and development of teachers of English as a foreign language, *Colombian Applied Linguistics Journal* 17 (2), 302–312.

Hone, K S and El Said, G R (2016) Exploring the factors affecting MOOC retention: a survey study, *Computers and Education* 98, 157–168.

Lowenthal, P, Snelson, C and Perkins, R (2018) Teaching Massive, Open, Online, Courses (MOOCs): tales from the front line, *International Review of Research in Open and Distributed Learning* 19 (3), 1–18.

Milligan, C and Littlejohn, A (2017) Why study on a MOOC? The motives of students and professionals, *International Review of Research in Open and Distributed Learning* 18 (2), 92–102.

Misra, P K (2018) MOOCs for Teacher Professional Development: Reflections, and Suggested Actions, *Open Praxis* 10 (1), 67–77.

Popham, W J (2009) Assessment literacy for teachers: faddish or fundamental?, Theory into Practice 48 (1), 4–11.
Reich, J (2014) MOOC completion and retention in the context of student intent, *EDUCAUSE Review*, available online: www.educause.edu/ero/article/mooc-completion-and-retention-context-student-intent
Stiggins, R J (2014) Improve assessment literacy outside of schools too, *Kappan* 96 (2), 67–72.
Stoynoff, S and Coombe, C (2012) Professional development in language assessment, in Coombe, C, Davidson, P, O'Sullivan, B and Stoynoff, S (Eds) *The Cambridge Guide to Language Assessment*, Cambridge: Cambridge University Press, 122–129.
Trinity College London (2022) *Assessment literacy*, available online: www.trinitycollege.com/qualifications/teaching-english/resources/assessment-literacy

10 Promoting learning about language assessment through a multilingual community of practice: ALTE as a case study

Graham Seed
Cambridge University Press & Assessment, UK

Waldemar Martyniuk
Jagiellonian University, Poland

Lorenzo Rocca
Società Dante Alighieri, Italy

This chapter provides a case study of the Association of Language Testers in Europe (ALTE). It explores how the association, as a community of language testing professionals, engages in promoting awareness of language assessment issues to a variety of stakeholders using different methods, and, in particular, strives to serve language professionals operating in languages other than English. The methods focused on include:

- The ALTE Quality Management System and accompanying documentation
- The courses and events it holds and the resources it creates
- Policy advocacy at national and supranational levels, specifically with regard to issues surrounding migration and integration

W tym rozdziale przedstawiono studium przypadku Stowarzyszenia ALTE (Association of Language Testers in Europe). Omówiono, w jaki sposób stowarzyszenie, jako społeczność profesjonalistów zajmujących się testami językowymi, angażuje się w promowanie świadomości

zagadnień związanych z ocenianiem sprawności językowej wśród różnych interesariuszy, a w szczególności stara się służyć specjalistom działającym w językach innych niż angielski. Zaangażowanie ALTE obejmuje:

- System Zarządzania Jakością (ALTE Quality Management System)
- Kursy, konferencje i otwarte publikacje
- Działania polityczne na szczeblu krajowym i ponadnarodowym, w szczególności w odniesieniu do kwestii związanych z migracją i integracją

Questo capitolo presenta un caso di studio dell'ALTE (Association of Language Testers in Europe). Il contributo intende illustrare come l'Associazione, in quanto comunità di professionisti del testing linguistico e attraverso un ampio ventaglio di strumenti, sia impegnata a promuovere consapevolezza in merito alle questioni legate alla valutazione linguistica, coinvolgendo i diversi attori in gioco e supportando in particolare coloro i quali operano in lingue diverse dall'inglese.

I metodi presentati includono:

- Sistema di gestione di qualità ALTE e documentazione accessoria
- Corsi, eventi e risorse ALTE
- Sostegno a livello nazionale e sovranazionale in relazione alle politiche linguistiche in materia di migrazione e integrazione

Introduction

The Association of Language Testers in Europe (ALTE) is a professional community of practice, promoting the fair and accurate assessment of language ability across Europe and beyond. Its recent new constitution sets out a core objective to 'advance the education of the public in general, and particularly among those with an interest in language assessment' (ALTE 2019). To this end, ALTE provides practical expertise in learning about assessing language, primarily to stakeholders such as test designers, item writers and others involved in test development, production and administration, both at the individual and organisational level; but also to those involved in education and immigration policy-making at national and international levels.

Fundamental to ALTE's mission is the concept of *sustaining diversity*, which recognises that different assessment contexts depend on various features including the domain, purpose, characteristics and needs of the test-taking population. However, arguably the principal feature of note within *sustaining diversity* is that of language. The contexts addressed by ALTE are notably in non-English testing situations, reflecting the association's

commitment to the promotion of multilingualism and plurilingual repertoires, especially through the 25 languages currently represented by its full members.[1] These range from the globally widely spoken (French, Spanish, Portuguese, etc.) through to the national languages of smaller countries (Estonian, Slovene, etc.) and then to regional and/or minority languages (Basque, Galician, etc.).

The association's objective to promote knowledge and raise standards in language assessment is operationalised to be achieved through a number of methods, and this case study chapter will provide an overview of some of these:

- the development of ALTE's Code of Practice and Principles of Good Practice documents, resulting in a Quality Management System, primarily aimed at language testing organisations
- courses, events and resources, primarily aimed at individual professionals who may or may not be connected to a language testing organisation
- policy advocacy relating to language assessment, most notably at the European level and particularly in the area of migration and social integration.

Included below are some considerations as to the effectiveness of these different methods. Insights were gained through a small-scale qualitative study carried out in December 2021. Responses to open-ended survey questions were received from 20 testing professionals working across 10 languages other than English who had engaged with the methods described. Half of the respondents had been active in ALTE for five years or less, and the other half for six years or more. In terms of respondents' involvement in language assessment in general, five stated their experience was 10 years or less; nine stated between 11 and 20 years, and six more than 20 years. Despite a small sample, a broad range of the amount of professional experience can be seen among the participants. The results of the survey have enabled an evaluation of the extent to which the methods help the development of language assessment knowledge as an area of professional competence, and comments from the respondents of this survey will be shared throughout the discussion below.

1 ALTE's Full Members are institutions that have undergone an ALTE quality audit of at least one of their qualifications. They are distinct from Associate Members, which are organisations with an interest and active involvement in language testing, but are not Full Members and the quality of their tests cannot be guaranteed. At the time of writing (June 2022), there were 110 Associate Member organisations, from different parts of the globe representing different languages, in addition to the 33 Full Member organisations.

Development of the Quality Management System

From the initial foundations of ALTE over 30 years ago, when a number of European testing organisations met together to pursue common interests, ALTE has had a desire to provide a forum for discussion and collaboration, and to establish common standards to address issues of quality and fairness in language testing. The development of its Code of Practice, iterations of the Principles of Good Practice documents and the introduction of the Quality Management System are the specific tools used to identify and promote a shared understanding of language assessment.

In its early days, a toolkit to support language test development was created, including the 1994 Code of Practice (ALTE 1994a), later translated into 22 other languages, which 'set out the responsibilities of ALTE members towards exam users and takers, and laid down minimum standards to be met by all members' exams' (ALTE 2020b:14). While the Code of Practice was an agreed definition of the minimum standards to be met in quality language assessment, a more detailed paper called the *Principles of Good Practice for ALTE Examinations* (ALTE 1994b, 2001) was also produced to give guidance on what could be considered good practice and how that could be implemented over time.

The members felt it was important to facilitate and continuously monitor the quality of implementation of the guidelines in testing organisations, in order to have positive impact on their practices. Seventeen 'minimum standards' were established[2], followed by a Quality Management System involving an auditing process to evaluate the extent that members implement practices to meet the standards, relevant to each testing context. Originally conceived as a peer audit system, where members review each other's tests, the system continues today with a professional cadre of auditors, overseen by an elected standing committee. A successful audit results in the award of a 'Q-Mark', a mark of quality for language tests, which is recognised by several bodies including the Swiss ministry for migration (Schweizerische Eidgenossenschaft Staatsekretariat für Migration (SEM) 2022). Regardless of the result of the audit, there are always recommendations and suggestions for continuous improvement, recognising even the most developed of testing bodies has something to learn. For more about the history of the development of the Quality Management System, see ALTE (2020b:68–83), and for more information about the Q-Mark system itself, see www.alte.org/Setting-Standards.

2 An eighteenth minimum standard is currently in development, aimed at test providers demonstrating measures in place to prevent test misuse. This seeks to improve assessment literacy in the specific area of test use consequences.

For those individuals in an organisation preparing for a Q-Mark audit, there are advantages in developing professional competences on a personal level as well as for the organisation itself. A test developer from Italy reported in our qualitative survey that they saw the benefits of going through a Q-Mark audit in helping to better understand the 'path towards the construction of valid tests' as well as considering the impact of the tests on the testing population. A respondent working with German language tests remarked that 'the "forced" self-reflection and documentation has helped to discover weaknesses and potential for improvement'. Remarkably, in preparing for a Q-Mark audit, an assessment professional from Romania noted how they 'actually rebuilt all the assessment system guided by the MSs [minimum standards]'. After this change, the respondent from Romania noticed how the candidates saw the relevance, validity and connection between the test and the target language use domain with greater clarity.

The current Principles of Good Practice document (ALTE 2020a) is in its third edition, revised and updated through collaboration and engagement of ALTE members. Furthermore, as Saville (2020:5) points out, the principles 'have been written to be accessible to a wide audience with varying degrees of professional expertise and resources at their disposal. In this respect, the approach taken contributes to "language assessment literacy" and helps ALTE members and other participants to understand assessment better in their own contexts.'

The document reflects the overall philosophy of ALTE and is intended as a coherent set of guidelines that all ALTE Members subscribe to when creating and delivering language tests. In the introductory section the importance of the ethical aspects of language assessment is highlighted with reference to the ALTE *Code of Practice* (ALTE 1994a), and the ILTA *Code of Ethics* (ILTA 2000), which identifies nine fundamental principles which 'draw upon moral philosophy and serve to guide good professional conduct', each elaborated on by a number of annotations which clarify the nature of the principles. The proposed principles of good practice are aimed at ensuring that examinations offered by ALTE members can be shown to meet explicit criteria in terms of the following five main qualities as outlined below:

> The concept of examination usefulness requires that, for any specific assessment situation, an appropriate balance must be achieved between the five main examination qualities: validity, reliability, impact, practicality and fairness. In addition, for ALTE members as providers of examinations to users around the world, quality of service is an important consideration. (ALTE 2020a:16)

It is underlined in the document that the individual examination qualities should not be considered separately but as a coherent set defining the overall usefulness of the examination (cf. Bachman and Palmer 1996).

During the ALTE 1st International Digital Symposium in April 2021, a panel on the use and impact of the ALTE Principles of Good Practice was organised to offer insights into ways in which these guidelines contribute to the improvement of quality and enhancement of professionalism in member organisations. Representatives of ALTE members as well as partner associations contributed to the panel by offering critical reflections and comments on the usefulness of the document. The panellists reported, among others, cases of successful inclusion of the document as a reference and study material in a university MA programme, improved assessment literacy and raised professional profiles at the institutional level through the extensive use of the document and the implementation of the principles, and the usefulness of the guidelines in fostering professional quality and intellectual empowerment of practitioners as well as in communication with policy makers (Martyniuk 2021).

During a plenary session organised by the Latin American Association for Language Testing and Assessment (LAALTA) at the XLV Conference of the Federation of the Argentinian English Teacher Associations (Federación Argentina de Asociaciones de Profesores de Inglés – FAAPI) in September 2021, the ALTE Principles of Good Practice 2020 were discussed in the Latin American context (Martyniuk, Saville and Villa Larenas 2021). Despite the European context of its origin the document was also considered useful in Latin America for informing policy makers and international exam users, as well as for offering students at Latin American universities a wider picture of the complexity of language testing.

Courses, events and resources

For the individual language testing professional, ALTE has provided various courses, conferences, events and resources throughout its history, in order to increase its assessment literacy programme. While some are targeted at individuals working in member organisations, many others are made available to any interested professional.

Courses

ALTE ran its first Introductory Course in Language Testing in 2005. Hosted in a different European location once a year, this week-long course focuses on the practical application of testing and assessment theory. As of 2022, due to increased demand as well as the practical implications of COVID-19 travel restrictions, the course happens online twice a year. Places on the course

are highly sought after, but a scholarship provides a free place to a suitable professional from a region or sector historically under-represented in the international language assessment community who is able to cascade their knowledge to others, thus promoting language assessment literacy beyond the traditional ALTE membership. Other courses have been run over the last 15 years on a variety of different language assessment-related subjects, including item writing, statistics and assessment of language skills, and of varying length from a day to a week.

Through the small-scale study described in the introduction, as well as through regular feedback received after each course, participants have reflected on the usefulness of what they learn and how they put this into practice. They have also noted the chance to consider new perspectives gained not only from the course but also with peers working with different languages, giving them increased knowledge, competence and also confidence. One respondent in the small-scale study from Lithuania mentioned that 'the course has given me a better understanding of the overall process of language test development. It also improved my contribution at work: I felt more confident to participate in the process of language test development.'

Conferences and events

ALTE's biannual meetings and conferences are a key way for individuals to learn from other professionals in the field: not only by taking part in workshops and lectures, but also by having the opportunity to network and talk informally with other participants. The larger international conferences, held every three years, also provide these opportunities on a greater scale. In 2021, due to COVID-19 restrictions, a Digital Symposium was held using a unique interactive online platform. In addition to these, other events such as the meeting of Special Interest Groups (SIGs) help foster a community of engagement and development of particular areas within language assessment. SIGs allow a more specialised consideration of issues in language assessment, and ALTE SIGs have taken up the issues of the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001), less widely tested languages, testing for specific purposes and others.

Through the small study carried out in December 2021, as well as through regular conference feedback forms, conference attendees note the benefits of shared support and contact, keeping up to date with current trends and research and seeing the experience of other organisations in tackling problems. One respondent to the survey also noted how attending events counteracts the disadvantage of working in relative isolation in a smaller language testing institution.

Resources

Over the last 30 years, ALTE's members have engaged in creating a number of useful resources for the benefit of the wider testing community as much as themselves. Most of these are now available as free downloads from ALTE's website, and include the Principles of Good Practice and information about the Quality Management System already discussed.

An early notable reference guide to aid assessment literacy was the *Multilingual Glossary of Language Testing Terms* (ALTE 1998), itself a volume within the same Studies in Language Testing series as the volume you are now reading. The glossary provides definitions of key terms in the field, translated from English into nine languages – in some cases where no term had previously existed. The process of compiling the glossary was an exercise in assessment literacy in itself as testers from across Europe came together to share knowledge and experience and develop a common and transparent discourse in multiple languages, which had not happened until that time.

A later significant publication was the *Manual for Language Test Development and Examining* (ALTE 2011), produced on behalf of the Language Policy Unit of the Council of Europe, and its supplement *Guidelines for the Development of Language for Specific Purposes Tests* (ALTE 2018). Users of these guides note their usefulness for training item writing teams, designing in-house workshops, and developing new exams for specific purposes such as nurses' occupational language, or for migration purposes.

ALTE's 'Can Do' Project was a seminal influence on the development of the CEFR (Council of Europe 2001) as it developed and validated a series of performance descriptors, setting out what learners can do in a language, across five levels. This resource was also made available in 11 languages other than English, once again emphasising the value of a multilingual provision of resources.

A striking conclusion of the language assessment literacy potential of ALTE's resources comes from a Welsh assessment developer commenting in our survey on the benefits of being part of the community creating the materials: 'I feel that participating in the development of some of these resources gives a sense of ownership, and I am therefore more likely to apply them to our tests.'

Policy advocacy

In order to achieve its charitable mission, ALTE has built a growing awareness of the necessity to avoid a standalone role. This is particularly true in relation to the need for improving an ongoing dialogue with

policy makers, especially with those related to the educational field. This means that ALTE aims to be proactive on the one hand in listening and understanding, and on the other hand in explaining and advocating, by taking into account both the macro level of decision makers (for example Ministries of Education), as well as the micro level of local community (for example teachers in a particular school). It implies a commitment by ALTE to look at the context when facilitating a constructive dialogue with both types of stakeholders.

ALTE's role is therefore to point out the importance of making sure language testers are engaged in the advocacy of trying to establish continuous interaction with those in charge of defining language curricula, syllabuses and teaching materials, at supranational and national levels. In the 2010s, ALTE's annual celebrations of the European Day of Languages each September involved events and papers presented to the European Parliament (for example ALTE 2020c), and liaison with the European Commission also proved valuable in advocating the need for quality language assessment. More concretely, member organisations from ALTE conducted the 2012 European Survey on Language Competences for the European Commission (European Commission, Directorate-General for Education, Youth, Sport and Culture 2012).

In addition, ALTE promotes engagement with those involved in defining and introducing laws related to language requirements and compulsory tests at the national level. This is particularly true with regard to the migration context, where ALTE has assisted in providing advice for the implementation of these requirements, calling out misuse of the CEFR (particularly in relation to the migration context), and to the presence of language requirements along the migrants' journey of integration into a host country (Council of Europe 2014).

Within the ALTE SIGs, the LAMI (Language Assessment for Migration and Integration) SIG was formed in 2002 as a platform for language testers to support their attempts of ensuring test fairness within the migration context. Several outcomes of the LAMI SIG are provided here, to exemplify the aim of increasing language assessment literacy among policy makers.

- The LAMI booklet *Language tests for access, integration and citizenship: an outline for policy makers* (ALTE 2016) represents a position paper available in three language versions where LAMI takes into account ethical and technical concerns to advise policy makers on such issues.
- A 'LAMI Forum' was held in 2017 on learning-oriented assessment for migrants with a panel of both teachers and policy makers, in order to facilitate the dialogue between these two stakeholder groups.

- The ambition of putting weight also at supranational level led LAMI to make a closer collaboration with LIAM (Linguistic Integration of Adult Migrants), a Council of Europe project launched in 2006 under the umbrella of the Educational Department.
- As an example of the coordinated action of ALTE and the Council of Europe, LAMI and LIAM conducted the most recent survey on language and knowledge of society requirements and learning opportunities for migrants among member states (Council of Europe–ALTE 2020). The survey was the largest in scope in comparison to the three previous ones; 41 European countries responded, allowing the formulation of evidence-based policy recommendations.
- As a result of the survey, ALTE–LAMI agreed on an action plan centred on highlighting the value of a profiling approach, which allows a representation of the different ability levels in different skills of a particular learner. This also led to the hypothesis of partial language certifications, especially in order to preserve the rights of the most vulnerable test-takers such as migrants with low literacy who can be awarded on speaking and listening skills only. LAMI is therefore currently working on increasing language assessment literacy by developing practical tools for teachers working with illiterate and low-literate learners.
- As an additional example of the cooperation with the Council of Europe, these tools are based on LASLLIAM (Literacy and Second Language Learning for LIAM), the Council of Europe's reference guide launched in June 2022 (Council of Europe 2022). The guide aims at sustaining the alignment between curriculum, teaching and assessment, supporting its recognition across Europe, and making the alignment accessible for the target learners involved in the simultaneous, interrelated and nonlinear processes of acquiring literacy and a second language.

Conclusion: Future challenges

The case study of ALTE as an engine of language assessment literacy is unique in that it caters to different stakeholders – primarily individual professionals, language testing organisations, and policy makers – through the different methods described above. In particular, ALTE has catered for those working with languages other than English and provides a community-style network to disseminate knowledge and competence.

Over its 30 years of existence, ALTE has both contributed and reacted to developments in language assessment, especially in Europe with the stakeholders it traditionally works with. The challenge for ALTE is to continue to advance thought leadership in defining quality language

assessment, for the benefit of both its increasing institutional and individual membership, and for the wider language testing community. This needs to take into account the social, technological, linguistic and educational changes that have come about over the last decade, and in particular since the beginning of the COVID-19 pandemic.

The future expansion in ALTE's activities could increase its reach to more people in those stakeholder groups; additionally, the inclusion of other stakeholders, such as teachers, test users and learners into its language assessment literacy programme is something to be considered, and work is already underway in collaboration with Eaquals, an international organisation promoting quality in language education, to meet the perceived needs of the language teaching community in this respect.

While multilingualism is very much at the forefront of ALTE's aims, the fact that its language of operation remains English in order to facilitate global communication is arguably both necessary and yet at odds with its aims. Providing courses in languages other than English as a mode of operation, together with more localised events for assessment literacy, are challenges to be met by ALTE in the coming years.

References

ALTE (1994a) *Code of Practice*, available online: alte.org/resources/Documents/THE%20ALTE%20CODE%20OF%20PRACTICE%201994.pdf

ALTE (1994b) *Principles of Good Practice for ALTE Examinations*, Internal discussion paper.

ALTE (1998) *Multilingual glossary of language testing terms*, Studies in Language Testing Volume 6, Cambridge: UCLES/Cambridge University Press, available online: www.alte.org/resources/Documents/9780521658775book_D.pdf

ALTE (2001) *Principles of Good Practice for ALTE Examinations – 2nd version*, Internal discussion paper.

ALTE (2011) *Manual for Language Test Development and Examining*, Strasbourg: Council of Europe, available online: www.alte.org/resources/Documents/ManualLanguageTest-Alte2011_EN.pdf

ALTE (2016) *Language tests for access, integration and citizenship: an outline for policy makers*, available online: alte.wildapricot.org/resources/Documents/LAMI%20Booklet%20EN.pdf

ALTE (2018) *Guidelines for the Development of Language for Specific Purposes Tests*, available online: www.alte.org/resources/Documents/6093%20LSP%20Supplement%20-%20WEB.pdf

ALTE (2019) *ALTE – Association of Language Testers in Europe. Association CIO Constitution*, available online: www.alte.org/resources/Documents/ALTE%20CIO%20Constitution%20Registered.pdf

ALTE (2020a) *Principles of Good Practice*, available online: www.alte.org/resources/Documents/ALTE%20Principles%20of%20Good%20Practice%20Online%20(Final).pdf

ALTE (2020b) *The History of ALTE. The Association of Language Testers in Europe: The first 30 years*, available online: www.alte.org/Our-history

ALTE (2020c) *Language Learning and Assessment in Europe. Briefing Paper to European Parliament*, available online: www.alte.org/resources/Documents/ALTE%20Briefing%20Paper%20to%20European%20Parliament.pdf

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

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

Council of Europe (2014) *Recommendation 2034. Integration tests: helping or hindering integration?*, available online: pace.coe.int/en/files/20482

Council of Europe (2022) *Reference guide on literacy and second language learning for the linguistic integration of adult migrants (LASLLIAM)*, Strasbourg: Council of Europe.

Council of Europe–ALTE (2020) *Linguistic integration of adult migrants: requirements and learning opportunities*, Strasbourg: Council of Europe.

European Commission, Directorate-General for Education, Youth, Sport and Culture (2012) *First European Survey on Language Competences: Final Report*, Brussels: European Commission Publications Office.

ILTA (2000) *ILTA Code of Ethics*, available online: www.iltaonline.com/page/CodeofEthics.

Martyniuk, W (2021) *The use and impact of the ALTE Principles of Good Practice*, panel presentation at 1st ALTE Digital Symposium, 28 April 2021.

Martyniuk, W, Saville, N and Villa Larenas, S (2021) *Defining the principles of good practice in validity and reliability in language assessment – how useful are ALTE's principles for the Latin American context?*, presentation at XLV FAAPI Conference 2021, 17 September 2021.

Saville, N (2020) Foreword, in ALTE (2020) *Principles of Good Practice*, Cambridge: ALTE, 5.

Schweizerische Eidgenossenschaft Staatsekretariat für Migration (SEM) (2022) *Liste der anerkannten Sprachzertifikate zum Nachweis der Sprachkompetenzen im Rahmen von ausländer- und bürgerrechtlichen Verfahren*, Köniz: Schweizerische Eidgenossenschaft Staatsekretariat für Migration.

11 Assessment literacy through the design of analytic scales

Joaquín M Cruz Trapero
Universidad de Jaén, Spain

In this chapter we describe how a group of Spanish policymakers helped improve the assessment competence of a community of language professionals through a project for developing a set of analytic scales for productive skills in language proficiency tests. We describe the:

- Context in which the project was developed
- Three stages of the project and the main challenges faced in each of them
- Six-step protocol used to design and validate the scales
- Perception of participants about the improvement of their assessment competence

Este capítulo describe cómo un grupo de gestores institucionales en el ámbito de la educación contribuyó a mejorar las competencias en evaluación de un conjunto de profesionales. Esta contribución giró en torno al desarrollo de diferentes escalas analíticas destinadas a la evaluación de destrezas productivas en exámenes de dominio. En el capítulo analizaremos:

- El contexto en el que el proyecto se desarrolló
- Las tres fases del proyecto y los principales retos a los que hubo que hacer frente
- El protocolo de seis pasos usado para desarrollar y validar las mencionadas escalas
- La percepción de los participantes sobre la mejora de sus competencias en evaluación

The context

Foreign language teaching and assessment in Spain is partially managed by a public network of 449 Official Language Schools (OLS) (Ministerio de Educación y Formación Profesional (MEFP) 2021). These schools provide evening and online lessons in multiple languages for students of different profiles, from high-school students to adults who seek to improve their language level to climb up the work ladder. They also prepare high-stakes proficiency tests for Levels A2 to C2 of the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001). These tests are the most widely recognized in Spain and their results are used nationwide to certify language levels in undergraduate and postgraduate studies, as well as in job recruitment processes. Every year, the OLS system caters for approximately 400,000 students (MEFP 2019:2), who find in it a quality alternative to private language teaching and certification, which has become a major industry in Spain since the publication of the CEFR (Cruz 2016b). Despite having a common curriculum for teaching and common specifications for test design (Boletín Oficial del Estado (BOE) 2017), the decentralisation of Spanish education allows for a great deal of diversity and autonomy across OLS. In practical terms this means that each one of the 17 Spanish regions develops its own curriculum for teaching and its own suite of proficiency tests.

Language professionals starting their careers in OLS are mostly very well trained to teach but may lack the necessary skills to design high-stakes tests. Even when they lack specific training in assessment, OLS professionals are often expected to join the regional commissions in charge of developing the proficiency tests and rating scales which will be later used across their regions. This may result in unpiloted tests and poorly constructed assessment scales. Although OLS tests and scales are aligned with the CEFR and, since its publication, with the Companion Volume (CV) (Council of Europe 2020), they are not always validated quantitatively.

Following the publication of the CV, Spanish legislation (BOE 2017) made it compulsory for OLS to include mediation in their proficiency tests. Mediation had been introduced earlier in the CEFR in 'the move away from the four skills, as one of the four modes of communication, namely reception, production, interaction and mediation' (Council of Europe 2020:35). Some regions took advantage of this change in Spanish legislation and decided to revisit their tests not only to provide for mediation but also to improve pre-existing versions of their scales for Production and Co-production of Written Texts (PCWT), and for Production and Co-production of Oral Texts (PCOT), the productive parts of OLS proficiency tests.

Even before this revision of scales occurred, the policy makers of one particular region, the Canary Islands, decided to train a group of language

professionals to develop a new set of analytic scales for the OLS of their region. The Department of Education and Universities (Consejería de Educación, Universidades, Cultura y Deportes, henceforth CEUCD) of the Canary Islands designed a three-stage project for in-house training of a group of OLS language professionals. The objective of the project was to provide them with the necessary competency in assessment to develop their regional scales, which later on would be used in all the OLS of the Canary Islands.

This paper describes how this project brought together the unrelated and seemingly inharmonious fields of teaching, assessment, linguistics and psychometrics (cf. Spolsky 1995), and the positive impact that it had on the whole community of OLS professionals.

The project

As mentioned above, the objective of the project was to provide language professionals from different OLS in the Canary Islands with the necessary competence to design and validate a new set of rating scales for their region. The scales in existence had been widely criticized by the OLS community since they were not properly designed. To achieve such a goal, the project was divided into three stages (see Figure 1): 1) consultation, 2) training, and 3) scale development. Throughout the project, a set of 10 analytic scales was designed, five apiece for PCWT and PCOT, each set for Levels A2 to C2 of the CEFR (CEUCD 2020, 2021).

	Stage 1 Consultation	Stage 2 Training	Stage 3 Scale Development
Participants	All OLS professionals in the region	16 OLS professionals	6 OLS professionals for scale development; 23 OLS raters for piloting; 100 OLS professionsals during the presentation of results
Tasks undertaken	Online survery	Training workshop, with initial theoretical lecture and practical work on the first step of the protocol for scale development	Full scale developmet; Piloting of scales; Final workshop to present the results to the OLS community
Data collected	Possible areas of improvement in the devlopment of scales	Structure of scales (type, number and structure of components, and number of bands)	Final version of the scales; Statistical analysis of scales

Stage 1: consultation

It was clear from the beginning that consensus would be essential for the success of the project. The assumption was that if all OLS professionals in the Canary Islands participated at a certain point in the design of the scales, their implementation would be perceived as a joint success rather than as an imposition. Thus, in the first stage of the project, all OLS professionals in the region were asked to do an online survey. The survey was developed by the regional Department of Education and Universities, the original promoter of the project, and it was sent to the 225 OLS professionals who, in November 2017, attended a workshop on measurement instruments for productive skills. The main objective of the survey was to analyse the experience of such professionals using former scales and to know their view on what aspects might be improved. The survey contained 20 close-ended questions on their preferences with regard to the number of bands that the new scales should have, and on the number and internal structure of its components (task adequacy, lexical and grammatical competence, fluency, etc.).

The answers obtained in the survey revealed some clear tendencies. For example, respondents concluded that the number of scale components should be reduced from seven to five, each one with five bands. Respondents preferred the new scales not to weight components and to identify 'plus' levels (A2, A2+; B1, B1+, etc.) (Council of Europe 2001:32). The survey also made clear that there was not much consensus on how the scale components should be chosen and defined.

Stage 2: training and early decisions

The second stage of the project was devoted to training. As we pointed out earlier, lack of training in assessment is an important concern among OLS professionals in Spain. Again, all language professionals in the Canary Islands were invited to participate in an 8-hour workshop which took place in December 2017. Sixteen language professionals were then chosen as representatives of the different schools and languages taught and assessed in the region.

The first half of the workshop contained a lecture on basic aspects of assessment, namely construct definition, validity, reliability, fairness and the difference between achievement and proficiency tests. Introducing these concepts was helpful in two ways. First, it helped to raise awareness of some concepts that were unfamiliar to several participants and, second, it established a common language which was very helpful throughout the project. The theoretical conceptualization of validity was particularly helpful since it provided a mental scheme for participants to base their work on. Following Bachman (1995) and Messick (1989), validity was presented as a unitary concept built through multiple evidences, and not as a group of

different isolated validities. In Messick's (1989:13) words, '(v)alidity is an integrated evaluative judgement of the degree in which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment' (italics in the original). Though not explicitly mentioned during the workshop, Wolfe and Smith (2007a, 2007b) was taken into account to operationalize Messick's view of validity, as they present clear and practical guidelines to investigate the different elements contained in Messick's (1989) view of validity.

During the second part of the workshop participants became familiar, through examples and practical exercises, with the different steps of the protocol that would be used in the design and validation of the new scales (Cruz 2016a). These steps are broken down in Figure 2.

The first step of the protocol took place during the second part of the workshop through group work and discussion. Most relevant decisions for scales are usually taken at this early stage, which requires intensive debate to create common understandings. This debate is best carried out face-to-face, which allows for immediate feedback and facilitates decision-making.

Figure 2 Protocol for the design of rating scales

Description of the step	Actions to be carried out
1. Discussion of previous considerations	a. Choose the type of scale: primary trait, holistic or analytic.
	b. Identify and describe the different scale components.
	c. Decide the number of bands.
	d. Consider the way in which scores will be presented.
2. Development of the descriptors	a. Select the CEFR tables that contain relevant descriptors.
	b. Distribute CEFR descriptors as anchor referents.
	c. Fill in intermediate and incomplete bands.
3. Qualitative validation	a. Consult other experts about the first version of the scale.
	b. Fine-tune the scale following the feedback from 3a.
4. Quantitative validation	a. Pilot the second version with at least 2 raters and 30 candidates.
	b. Analyze the results obtained in 4a through Rasch.
	c. Fine-tune the second version following the feedback from 4b.
5. Implementaion	a. Prepare the third version in a user-friendly format.
	b. Conduct rater-training and benchmark sessions.
	c. Use the scale in live tests and gather data.
	d. If possible, repeat 4. If not, analyze inter-rater reliability.
6. Revision	a. Set up a cycle of revision.
	b. Gather data from different live administrations to draw conclusions.
	c. Fine-tune the scale if necessary and repeat stage 5.

The debate on the number of scale components and bands, for example, was particularly productive. The consensus was that as the number of components and bands increases, so does the cognitive load for raters. Although at the initial stage of consultation most respondents advocated for scales with five components and five bands, this was only true of PCOT; PCWT scales were designed with four components and five bands. One extra component was necessary in PCOT to account for aspects inherent to oral production such as phonetics and fluency.

Identifying and describing scale components was particularly challenging since scale developers had to harmonize their individual conceptualization of language with the construct of the test, and also had to dissociate their role as teachers from their role as scale and test developers. In other similar projects we have found that these problems are particularly frequent among teachers who, if not trained properly, tend to design and mark proficiency tests as if these were achievement tests designed for the students in their lessons. They may rate tests impressionistically or by comparing one candidate against another, and not against the rating scales, which opens the door to halo effect and bias.

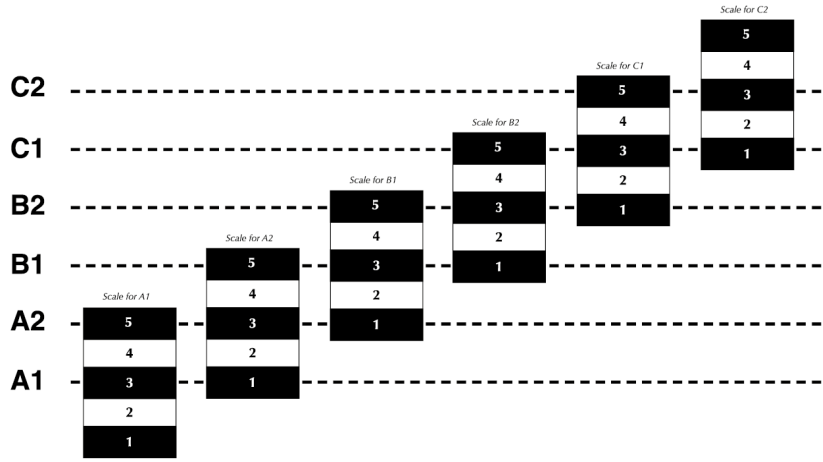
The debate on bands also proved enriching. Following the recommendations of the CEFR (Council of Europe 2001:181–182), Band 3 of the scales was considered the 'pass' level. Therefore, in a B2 scale, for example, B2 descriptors were used for Band 3 and the formulation for descriptors in Bands 1 and 5 was drawn from adjacent levels, B1 and C1 respectively. When deciding on Bands 2 and 4, some team members advocated for leaving them blank, while others insisted on their being defined in full. Defining all bands yields more comprehensive scales while it also increases the cognitive load on raters – the same issue addressed in the selection of components. Leaving some bands blank to identify candidates who share characteristics of adjacent bands eases the use of the scale but, as we will see in the section 'Stage 3: Scale development', at the risk of potential misinterpretations of its use. A full description of Bands 2 and 4 was also troublesome because, according to the design model (see Figure 2), these correspond to 'plus' levels, which are either not evenly described in the CEFR or not described at all.

Thanks to this design, for example, Bands 3, 4 and 5 in the B1 scale can be used as Bands 1, 2 and 3 in the B2 scale, etc. The model, however, posed some complications, such as the definition of Band 5 in C2 scales, for which there are no descriptors in the CEFR (for further detail on how to solve this see North (2000, 2020) and PETRA-E Network (2016)).

Stage 3: scale development

Out of the 16 participants in Stage 2, six volunteered to move on to the next stage, thus becoming the actual scale developers. Experience shows that larger numbers of scale developers may slow down the process.

Figure 3 Scalable design model



Steps 2 to 6 of the protocol took place during Stage 3 of the project. At Step 2, scale developers compiled the relevant descriptors from the CEFR, taking into account the first drafts of the CV, as well as the regional curriculum of the Canaries. After the descriptors were compiled, the resulting matrix was carefully edited in one spreadsheet. Since descriptors are not described with the same level of detail across all CEFR tables, the matrix was incomplete at some points and redundant at others. In several online meetings held by the six scale developers, descriptors were reduced or (re)written when necessary. The CEFR and curriculum tables they came from were noted down, which guaranteed their traceability.

In Step 3 of the protocol, after the scales had been finished and properly formatted, 23 OLS professionals other than the scale developers were consulted for qualitative validation. They checked for over- and underdefined descriptors, for lengthy or wordy formulations, and for typos. This helped fine-tune the descriptors.

For Step 4, the scales were piloted statistically. The statistical results collected at this stage would be a clear indicator of the extent to which a good scale could be developed by the team whose training we have described above.

The framework chosen for the statistical analysis was Linacre (1999), which proposes a series of guidelines for analysing scales. These are: 1) there must be at least 10 observations of each band (Linacre uses the term 'category' instead of 'band'); 2) there must be regular observation distribution; 3) average measures must advance monotonically from band to band; 4) outfit mean-squares must be less than 2.0; 5) step calibrations

Table 1 Main statistics of rating scale utility

Scale	Band	Production and co-production of written texts				Production and co-production of oral texts			
		Observations	Average measure	Outfit mean squares	Step calibration	Observations	Average measure	Outfit mean squares	Step calibration
A2	5	105	1.01	1.0	.99	70	1.63	.8	1.43
	4	134	.47	.9	.82	96	.63	.9	1.08
	3	262	–.09	1.1	–.74	228	–.29	1.0	–.96
	2	180	–.64	.9	–1.07	198	–1.43	1.2	–1.55
	1	159	–1.34	1.0	–	348	–2.96	1.0	–
B1	5	106	1.22	.9	1.25	62	.56	1.0	1.26
	4	149	.73	.9	.92	152	.20	1.0	1.11
	3	266	–.09	1.0	–.58	450	–.15	1.0	–.83
	2	200	–.60	1.1	–1.60	269	–.50	1.0	–1.54
	1	119	–1.61	1.1	–	117	–.91	1.0	–
B2	5	61	1.03	1.0	1.41	22	.90	1.2	1.52
	4	112	.64	.8	.74	53	.28	.9	1.40
	3	219	–.40	1.2	–.61	284	–.79	1.0	–1.03
	2	229	–.95	.7	–1.54	333	–1.53	.8	–1.52
	1	219	–2.11	1.2	–	358	–2.37	1.0	–
C1	5	64	1.67	.9	1.79	20	.45	1.5	1.39
	4	135	.49	.9	.33	50	.05	1.0	1.15
	3	171	–.35	1.1	–.62	243	–.80	1.0	–1.19
	2	191	–1.14	1.0	–1.49	231	–1.31	.7	–1.36
	1	262	–2.29	1.0	–	336	–2.10	1.0	–
C2	5	40	1.21	1.0	1.72	32	.71	1.3	1.62
	4	99	.45	.9	.54	92	.31	.8	.84
	3	179	–.60	1.3	–.62	245	–.48	.9	–.83
	2	239	–1.21	.8	–1.64	258	–1.28	.9	–1.63
	1	283	–2.35	.9	–	268	–2.08	1.1	–

must advance; 6) ratings must imply measures, and measures must imply ratings; 7) step difficulties must advance by at least 1 logit; and 8) by less than 5.0 logits.

For this purpose, after a training session, 25 raters from different OLS of the region analysed 30 samples per level (A2–C2) and scale (PCWT and PCOT). Again, the intention was to involve as many OLS professionals as possible while helping them improve their assessment competence. Table 1 summarizes the main results obtained after the analysis of the scales, which was carried out with Facets (Linacre 2014).

The number of observations obtained (Guideline 1) provided sufficient information about the mechanics of the scales. These observations displayed a regular unimodal distribution (Guideline 2) which pivoted

around intermediate or lower bands. The monotonical advance of average measures (Guideline 3) points to the average ability of candidates scoring at a particular band. According to Linacre (1999:111), 'observations in higher categories must be produced by higher measures' or higher abilities. As we can see in the Average measure columns, the bands of the different scales actually separated different levels of competence, with lower bands linked to lower candidate abilities and upper bands linked to more able candidates (for example, in the PCWT A2 scale, the average measure of candidates in Band 5 is 1.01, which is higher than the average ability of candidates in Band 1, -1.34). The outfit mean square of bands (Guideline 4), an indicator of randomness and overpredictability, was within the expected range. The advance of step calibrations (Guideline 5), the meaning of measures (Guideline 6), the advance of step difficulty (Guideline 6) and the logits that separate such steps (Guidelines 7 and 8) concern the scale's inferential value. Small separations suggest that two adjacent bands might be merged. Big separations indicate that bands represent too wide a range of performance, which generates 'dead zones' in the middle of a band thus leading to a loss of precision (cf. Linacre, 1999:119). As seen in the columns for step calibrations, though steps advanced, some bands displayed short separations, with irregularities in blank bands. As subsequent statistical analyses of the same scales proved, this was due to misinterpretations of the meaning and use of blank bands.

Although there was room for improvement in the design of the scales (specifically as regards the mechanics of use of blank bands), these results arguably provided the best indicator of the progress in the assessment literacy of OLS language professionals: without any previous experience in scale design (see the section 'Discussion'), thanks to a training workshop and to the guidance of a design protocol (Cruz 2016a), they were able to create a valid measurement tool *ex novo*. Such improvement is reinforced by the perception of participants also reviewed in the 'Discussion' section.

For Step 5 of the protocol, a second online workshop was organized. Approximately 100 participants attended this workshop, in which the whole design process was reported, paying special attention to the results of the statistical validation. The statistical data were adequately formatted and explained for the purpose of presenting the main findings to support the consistency of the new scales. The scales were used for the first time in the official test sittings of 2020 to assess 19,850 candidates according to data provided by the local administration (see CEUCD 2020, 2021).

Discussion

Approximately 270 OLS professionals (roughly 75% of the whole community of OLS professionals in the Canaries) participated in the project at different stages:

- 225 participants in the first survey
- 16 trainees in the workshop
- 6 scale developers
- 23 OLS language professionals during the qualitative validation
- 25 raters trained for the quantitative validation
- 100 attendees to the last online workshop

These data demonstrate that most OLS professionals in the Canary Islands were involved in the project at some point. Yet, despite this high level of participation, one question remained unanswered: did the project actually have any positive impact on the participants?

To answer this question, we developed a second survey in 2022, after the scales had been finished and were already undergoing their first revision cycle (Step 6). We pooled the opinion of the six OLS professionals who had been present at all stages of the project, and who also had management positions at the local government level or in their own OLS, which gave them a privileged perspective regarding the impact of the project. This survey, validated by three assessment experts, contained dichotomic questions on previous training in assessment, open-ended questions and Likert scales on the impact that the scales had had on the OLS community of the region. All respondents answered the survey as indicated in its instructions.

As regards the individual competence of the participants, the survey confirmed that none of them had received any specific training in assessment at university. They claimed that most of their training in assessment prior to the project had come from workshops organized by the Department of Education and Universities of the Canary Islands. All of them rated the content of the workshop with five points out of five. When asked about their assessment competencies, they noted that the project had contributed to their professional development in the following areas:

- General assessment as described by the CEFR
- Language proficiency assessment
- Assessment through scales

All of them declared that their experience in the project had helped them improve as professionals insofar as they were now able to separate their roles as teachers from their roles as language proficiency raters. Respondents reported having learnt to make both tests and judgements on scores fairer

by not rating candidates impressionistically. They also reported a growing personal interest in assessment and in statistics applied to quantitative validation. Some of them even trained other colleagues in-house on the use of the scales.

Besides this individual improvement of the assessment competence of participants, the survey also provided interesting insights into the washback effect (Davies et al 1999) that the new scales had on the whole community of OLS in the Canary Islands. Respondents reported that some groups of co-workers had been reluctant to accept the new scales after their release, perhaps due to the fact that they felt more comfortable with the previous procedures. Nevertheless, the general perception was that scales had a positive impact on fairness in tests and that they had encouraged all OLS professionals to reflect on their assessment competence.

Final remarks

We have presented the different stages of a project whose main objective was to develop a fair and sophisticated assessment tool by training language professionals who had no prior experience in scale development. The stable results of the reliability analysis of the scales and the improvement of the assessment competence of the participants as perceived by themselves prove that the project succeeded in its main objective. In fact, similar projects carried out with other local administrations had similar positive results (not reported here). Experience tells us that theoretical discussions on basic aspects such as construct definition, validity, reliability or fairness are paramount for participants at early stages of assessment projects, particularly when they lack previous experience in proficiency assessment or when they come from the field of language teaching or linguistics. Although language proficiency assessment builds on linguistics and can benefit from the experience of language teachers, the principles and techniques governing these three disciplines are not always interchangeable.

The results presented here show that despite the apparent complexity of this type of project, training in assessment may serve a double objective: creating valid assessment tools and improving the assessment skills of a whole community of language professionals which, in turn, may have a positive effect on fairness for candidates.

We would like to encourage Spanish universities to re-evaluate the importance of language proficiency assessment in their curricula, particularly at a time when testing has become an industry of unprecedented impact. We would like to encourage all OLS in Spain to revisit their current assessment methods and take advantage of the challenge posed by the assessment of mediation. There is no doubt that both universities and OLS have the necessary resources and a committed

community of professionals able to transform foreign language teaching and assessment in Spain.

References

Bachman, L (1995) *Fundamental Considerations in Language Testing*, Oxford: Oxford University Press.

BOE (2017) Real Decreto 1041/2017, de 22 de diciembre, por el que se fijan las exigencias mínimas del nivel básico a efectos de certificación, se establece el currículo básico de los niveles Intermedio B1, Intermedio B2, Avanzado C1, y Avanzado C2, de las Enseñanzas de idiomas de régimen especial, *BOE* 311, available online: www.boe.es/eli/es/rd/2017/12/22/1041/con

CEUCD (2020) Resolución de 17 de diciembre de 2020, por la que se convocan pruebas para la obtención de la certificación de enseñanzas de idiomas de régimen especial destinadas a la población escolar de centros docentes públicos de la Comunidad Autónoma de Canarias, y se dictan instrucciones para su organización y aplicación, *BOC* 334, available online: www.boe.es/eli/es/res/2020/12/17/(2)

CEUCD (2021) Resolución de 1 de diciembre de 2021, por la que se convocan pruebas para la obtención de la certificación de enseñanzas de idiomas de régimen especial destinadas a la población escolar de centros docentes públicos de la Comunidad Autónoma de Canarias, y se dictan instrucciones para su organización y aplicación en el curso 2021–2022, *BOC* 252.

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

Council of Europe (2020) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume*, Strasbourg: Council of Europe Publishing.

Cruz, J (2016a) *A Protocol to Design a CEFR-linked Proficiency Rating Scale for Oral Production and its App Implementation*, Jaén: Universidad de Jaén.

Cruz, J (2016b) The industry of testing and mutual recognition in Southern Europe, *The Grove: Working Papers on English Studies* 23, 11–22.

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.

Linacre, J M (1999) Investigating rating scale category utility, *Journal of Outcome Measurement* 3 (2), 103–122.

Linacre, J M (2014) *Facets Rasch Measurement Computer Program*, Chicago: winsteps.com, version 3.71.4.

MEFP (2019) *Estadística de las Enseñanzas no universitarias. Enseñanza de Lenguas Extranjeras. Curso 2017–2018*, available online: bit.ly/3ydwDPT

MEFP (2021) *Escuelas Oficiales de idiomas en España – Curso 2020–2021*, available online: bit.ly/3Hpzx8x>

Messick, S (1989) Validity, in Linn, R (Ed) *Educational Measurement*, New York: Macmillan, 13–103.

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

North, B (2020) *The CEFR Illustrative Descriptors: Validation Reference Paper for Researchers*, Strasbourg: Council of Europe.

PETRA-E Network (2016) *PETRA-E Framework of Reference for the Education and Training of Literary Translators*, available online: ecspm.org/wp-content/uploads/2016/04/PETRA-E.pdf

Spolsky, B (1995) *Measured Words*, Oxford: Oxford University Press.

Wolfe, E and Smith, E V (2007a) Instrument Development Tools and Activities for Measure Validation Using Rasch Models: Part 1 – Instrument Development Tools, *Journal of Applied Measurement* 8 (1), 97–123.

Wolfe, E and Smith, E V (2007b) Instrument Development Tools and Activities for Measure Validation Using Rasch Models: Part 2 – Validation Activities, *Journal of Applied Measurement* 8 (2), 204–234.

12 Using Lesson Study to develop teacher educators' language assessment literacy

Magnus Coney
Cambridge University Press & Assessment, UK

Ben Naismith
Duolingo, USA

This case study applied the Lesson Study approach for the professional development of a group of three language teacher trainers working on the Cambridge CELTA (Certificate in Teaching English to Speakers of Other Languages), a pre-service training course for teachers of English. It focuses on:

- Language assessment literacy and its importance for teacher training
- The Lesson Study approach as a tool for professional development
- The professional development of the trainers viewed through the lens of the Extended Interconnected Model of Professional Growth
- The trainers' views of the Lesson Study approach

Language assessment literacy (LAL) is recognised as an important aspect of teaching competency which greatly impacts the quality of student learning (Stabler-Havener 2018, Weideman 2019). However, there is little research into how LAL might be developed for stakeholders involved in second language (L2) teacher education (Villa Larenas 2020). A better understanding of LAL development for this group, i.e., teacher educators, could potentially lead to more effective instruction about language assessment for the L2 teachers in training (hereafter *trainees*).

In this case study we explore one form of collaborative professional development, the Lesson Study (LS) approach[1] of teacher-led research.

1 Lesson Study has been variously described as a model, a cycle, an approach (e.g., Coenders and Verhoef 2018), a framework, and a process (e.g., Fujii 2016). For consistency, this paper will use the broader term 'approach'.

Using a qualitative design, we followed the journey of three experienced teacher educators (hereafter *trainers*) as they planned, delivered, and reflected on an assessment training session for the Cambridge CELTA (Certificate in Teaching English to Speakers of Other Languages). Throughout this process, data were collected in various forms including observations in the researcher logs, trainers' written reflections, and interviews. By triangulating these data from different sources, we examined the viability of LS as a tool for developing the LAL of trainers (and, by extension, their trainees).

Background

Language assessment literacy of teacher educators

Language assessment literacy (LAL) is a broad term which encompasses an array of skills and knowledge related to assessment concepts, as well as the ability to apply these concepts in professional contexts (Deygers and Malone 2019). Here, the context of interest is entry-level teacher training courses in which trainers deliver sessions about assessment.

At present, there is a consensus that LAL is an important component of teacher education because it allows teachers to use, design, and evaluate assessment options appropriately (Weideman 2019). It is necessary to therefore consider the place of assessment literacy training in entry-level teacher training courses, including the LAL of the trainers themselves. After all, teacher education impacts teacher beliefs, which in turn drive pedagogical classroom decisions (Borg 2003). If trainers themselves lack LAL, or are uncomfortable teaching LAL, there may well be a knock-on effect. However, there is little research investigating the LAL of trainers. One study by Villa Larenas (2020) focused on the LAL of trainers in Latin America and found that they had not received specific training in language assessment. More generally, teacher education programs in Latin American higher education typically did not include specific assessment-related courses as part of their offerings.

On a global level, CELTA is one of the most widely recognised entry-level language teaching qualifications, with over 12,000 candidates annually in 48 countries (Cambridge Assessment English 2021a). The widespread take-up of the qualification around the world means that CELTA trainers are diverse in terms of their professional training and backgrounds. The requirements for becoming a trainer include substantial ELT experience and a recognised advanced qualification such as an MA TESOL or Cambridge DELTA. However, because these requirements are only broadly defined, the exact LAL training that CELTA trainers have received varies considerably.

On CELTA courses themselves, the syllabus section most directly relevant to assessment literacy is *The monitoring and evaluation of learning* (Cambridge Assessment English 2021b:11) covering the following syllabus points:

a. monitor learner behaviours in class time and respond appropriately
b. incorporate into their lessons some basic assessment procedures
c. make planning decisions on the basis of assessment.

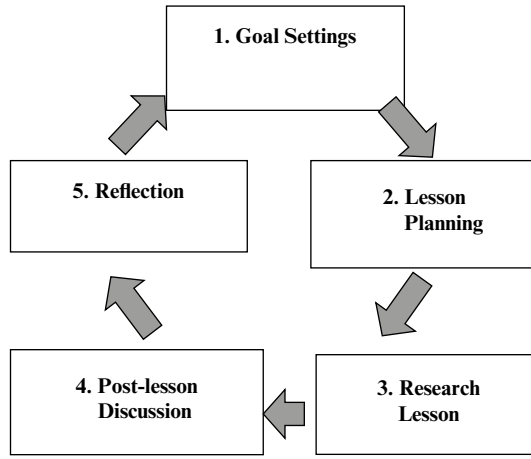
Because these points are themselves quite broad, trainers have a great deal of freedom to interpret them in different ways. There are no sets of prescribed teaching materials, and trainers have the autonomy to use whatever tasks and activities they deem necessary to facilitate trainee learning within the constraints of the course timetable. Consequently, the depth of training on assessment topics can vary widely across courses, with individual trainers prioritising LAL to varying degrees. To date, no studies have investigated these interpretations of syllabus points, i.e., what assessment elements trainers choose to include, exclude, or prioritise.

Lesson Study approach

Lesson Study is an approach to teacher development originating in Japan (Coenders and Verhoef 2018). The two main phases of LS are the *development phase*, where participants investigate an inquiry question and plan a lesson, and the *class enactment phase*, where a participant is observed teaching the lesson, then revises the lesson plan based on their reflections (Coenders and Verhoef 2018). In a complete LS cycle, the two phases can be further broken down into sub-phases, often differing in number and name depending on the author and context. Here, we follow the process as outlined by Fujii (2016:412) consisting of five sub-phases (see Figure 1). The development phase (Steps 1 to 2 below) begins with the identification of goals and inquiry questions, followed by the lesson planning stage. The class enactment phase (Steps 3 to 5 below) consists of the research lesson itself (where the participants investigate their own inquiry questions), a post-lesson discussion, and a final reflection. A key characteristic of LS is that it is participant-led, and as such, little external guidance is provided to participants.

A range of studies have previously revealed LS to be beneficial for teachers more generally (e.g., Cajkler and Wood 2016, Coenders and Verhoef 2018, Nami, Marandi and Sotoudehnama 2016). A review by Godfrey, Seleznyov, Anders, Wollaston and Barrera-Pedemonte (2019) found a variety of approaches to measuring the impact of LS. The review highlighted the importance of 'evidence of transfer of learning from lesson study' to the general classroom practice of teachers (2019:241). To our knowledge, LS has not yet been researched as a tool for the professional development of

Figure 1 The Lesson Study approach (Fujii 2016:412)



teacher trainers. The distinction between language teachers and teacher trainers is an important one because both the skills and knowledge they require meaningfully differs. For teachers, the focus is knowledge of language and how to teach it, whereas for trainers, the focus is knowledge of pedagogy and how to facilitate this development in teachers. As a result of such differences, it cannot be assumed that the impact of LS on language teachers and teacher trainers is comparable.

The Extended Interconnected Model of Professional Growth

One lens through which to visualise learning in LS is the Extended Interconnected Model of Professional Growth (EIMPG, Clarke and Hollingsworth 2002), adopted by Coenders and Terlouw (2015). The EIMPG was originally developed to describe teacher professional growth. However, it has also been adapted to analyse trainer learning (Perry and Boylan 2018), enabling researchers to consider a range of factors affecting professional development. In this model, the trainer's LAL is considered part of their *Personal Domain*. The knowledge, beliefs, and attitudes represented in this domain are increased through the interaction between three other domains: the *External Domain* (external sources of input), the *Domain of Practice* (the trainer's training activities), and the *Domain of Consequence* (the trainee learning outcomes).

The current study therefore seeks to use the EIMPG, especially the Personal Domain and Domain of Practice, to further explore the potential of the LS approach as a professional development tool for teacher trainers.

Specifically, we asked the following research questions (RQs) about the benefits of LS:

RQ1: To what extent do the trainers perceive Lesson Study to be an effective approach to professional development?

RQ2: How does the trainers' adoption of the Lesson Study approach lead to perceived professional development in terms of their language assessment literacy specifically?

Methodology

The trainers involved in this study were specifically invited by the authors to ensure participants with sufficient motivation, given the significant time commitment. All three trainers are highly experienced (10+ years of teaching and 8+ years of training experience), located in Canada, Germany, and the UK, respectively.

Following the LS approach described in the section above, the process was separated into five phases (Table 1). The trainers began the *goal setting* stage by meeting one another virtually (during which researcher logs were compiled) to decide on an area of LAL they wanted to develop, for example peer assessment or the principles of assessment. At this point, trainers were provided with the lesson focus for an introductory training input session on assessment and the relevant section of the CELTA syllabus. From this meeting, the trainers produced two inquiry questions. That the trainers determined the research questions is an integral component of the LS

Table 1 Lesson Study stages, procedures and collected data

Lesson Study stage	Procedures	Data collected
Goal setting	• One Zoom meeting • Select area of LAL to develop • Create inquiry questions • Independent research	• Researcher observations • Reflective journal entries
Lesson planning	• Two Zoom meetings • Independent material design	• Researcher observations • Reflective journal entries
Research lesson	• Research lesson delivery • Peer observations	• Researcher observations • Reflective journal entries • Lesson recording
Post-lesson discussion	• Two Zoom meetings	• Researcher observations • Reflective journal entries
Reflection	• Individual written self-reflection • Interviews with researchers	• Interview recordings • Reflective journal entries • Trainee end-of-course feedback

approach, reflecting the participant-led nature of this type of action research. Specifically, the participants were interested in how the trainees would respond to the content of the input session, and secondly, to what extent their own perceptions matched those of the trainees regarding the importance of the assessment topics in the session. They then spent time individually researching the area and were provided with three relevant assessment-focused articles to prompt their upcoming discussion and ensure shared understanding of the topic.

For the *lesson planning* stage of LS, two further meetings were held to collaborate on the planning of the research lesson, with trainers creating some of the materials independently in between meetings. The *research lesson* was then delivered online via Zoom by one trainer, enabling it to be recorded and observed by the other trainers (both synchronously and asynchronously due to time zone differences). The attendees were seven CELTA trainees on a course in Germany, diverse in age, first languages, and previous teaching experience. For the *post-lesson* discussion, trainers met for two Zoom sessions before completing the final *reflection* stage by completing their journal entries and conducting individual interviews with the researchers. During all these stages, researcher logs were further updated, for example with notes on trainer attitudes and their relative contributions to discussions.

A qualitative research design was chosen to enable close examination of the potential of LS. The primary instruments used were reflective journals, completed after each stage by each of the trainers, and a final semi-structured interview with one of the authors, consisting of 10 questions plus follow-ups to gather overall reflections and to build on the reflective journals. Finally, the researchers observed the meetings and research lesson and made notes in their logs. The journal entries, researcher logs, and interviews (transcribed using https://otter.ai) provided extensive data for qualitative analysis. The transcribed data were coded according to the EIMPG domains (described at the end of the previous section), and trainer comments were categorised according to the stages of the LS approach. The authors checked the suitability of the coding categories by independently analysing the data from one participant and discussing the results. The observational data from the meetings and research lesson, as recorded in the researcher logs, were compared to what the trainers reported in their journals and interviews in order to identify overlap or discrepancies.

Findings/Discussion

We now consider our findings in relation to the two RQs.

RQ1

Overall, there was unanimous consent that LS was perceived to be an effective method for stimulating trainer professional development. With respect to the specific sub-stages of LS (see the section 'Lesson Study approach'), the following trends emerged:

Goal setting/Individual research stage

All three trainers felt this initial stage to be an important and useful one. In particular, the resources provided by the researchers were appreciated, both for their contents and as a 'point of departure' (Trainer 2, hereafter T2) to structure the initial discussion. In fact, it was noted by two trainers that the resources did not particularly introduce any new knowledge about assessment, but rather, they were a good reminder of assessment basics and helped to stimulate reflection. The one challenge of this stage was finding time in their busy work schedules to complete the independent research in preparation for the goal setting meeting.

Collaborative lesson planning

The collaborative lesson planning took place over two hour-long sessions. The sessions started with discussion of *what* to include before moving on to *how* to do it. Key elements that the trainers wanted to include were stages for personal reflection, clarification of key assessment principles and types of assessment, and discussion of relevant assessment case studies.

This element of collaboration was also prominently featured in the trainers' journals and interviews, with the trainers finding this stage to be particularly useful and enjoyable. As T3 noted, 'I think probably the most useful thing was being able to bounce ideas off each other', a sentiment echoed by the two others. This collaborative element was perhaps particularly salient to the trainers because the act of team planning is so rare in their professional contexts; CELTA trainers often plan independently and then deliver training sessions in which they are considered the expert. The researcher logs confirm the consistent and lively collaboration between all three trainers.

Nevertheless, this stage was the most challenging for the group because it required careful managing of discourse given the different communication styles of the trainers. For example, one trainer could be seen to be the most effusive in the group sessions, but the most concise in their written reflections, whereas another displayed the opposite preference, writing extensively in their journal but speaking less in the group setting. A second challenge was coming to a consensus and finalising task designs and the task sequence:

> I realized when we were doing the planning together, that we left lots of things very kind of fluffy and open and not very concrete. (T3)

In part, this challenge may have been due to the mode of communication (Zoom) and because the trainers had never previously worked together. These challenges perhaps point to the need to establish basic ground rules to acknowledge different styles and mitigate risk.

Research lesson

The research lesson was delivered online in the final week of the CELTA course. In describing the lesson, all three trainers noted the trainee engagement, the appropriacy of the contents, and the quality of the delivery. These impressions were mirrored in the researcher logs. In addition, this perception seems to have been shared by the trainees themselves; in the post-course evaluation of the course's input sessions, they made numerous comments:

- 'The testing and exam session was particularly useful in underscoring the purpose and relevance of assessment.'
- 'I also found [T1]'s session on assessment very useful – the content was great, and really consolidated how useful monitoring and collaborative syllabuses can be.'
- 'The last online meeting about testing and exams was really useful! Although it would have been nice if it had come a bit earlier.'

Post-lesson discussion

The post-lesson discussion stretched over two sessions. These sessions focused on the perceived effectiveness of the different stages and tasks, potential alternatives to be used in the future, and the way in which the trainees reacted to and engaged with the session contents.

Many positive feelings were expressed regarding this stage, with trainers noting that the group reflection was enjoyable, engaging and gratifying. What was apparent was that the ideas discussed extended beyond the individual reflections, with new suggestions emerging through their exchanges. As T2 eloquently describes, 'a dialogic approach to reflection is always rich in potential for new insights'.

Reflection

Reflection was an integral component of the LS approach, with the trainers completing reflective journal entries after each stage. However, the final interviews with the researchers served as a *de facto* final summative reflection revealing trainers' holistic impressions of the LS approach and the success of the experiment. Four main conclusions emerged:

1. The tangible outcome (a planned CELTA assessment session) was of great value as all the trainers would use it themselves in the future (with minor changes).
2. The LS approach as a whole was coherent and cohesive, and all of the stages were useful in different ways.
3. The original inquiry questions were not always at the forefront throughout the process, and in the end, only the first inquiry question could be fully answered. The trainees responded very positively to the input session and found it to be useful; however, it is unclear to what extent the trainers' and trainees' perceptions aligned.
4. The greatest challenge for LS in general was timing. Satisfactorily completing the entire sequence inherently required a significant time commitment, and the planning stage in particular was demanding in this regard.

RQ2

The qualitative data were analysed using the EIMPG to understand and categorise the changes experienced by the trainers. For this study, certain domains of the EIMPG are more clearly supported by the data, specifically the Personal Domain and the Domain of Practice, while changes to the Domain of Consequence and the External Domain are only indirectly evidenced.

Personal Domain

The Personal Domain of the EIMPG encompasses trainers' knowledge and beliefs. Here, all three trainers reported some changes in their understanding of assessment, mainly in terms of gaining a broader view or a consolidation of beliefs they already held, rather than acquiring new knowledge. They also described changes in the way they thought about assessment relating to CELTA trainees:

- I found myself thinking about the outlined principles … and the various types of assessment done on CELTA. (T1)
- [It] helped me focus on how assessment literacy can be highlighted better for CELTA candidates. (T2)
- It's really made me think about my own assessment practices now as a trainer. (T3)

The value the trainers attached to thinking about assessment as it related to their CELTA experience may also have affected their planning of the research lesson, where they emphasised the importance of trainees thinking about assessment in relation to their own experiences (discussed further below).

Domain of Practice

Secondly, we consider the Domain of Practice, i.e., trainers' professional practice. This domain requires evidence of observed changes, meaning that through their delivery of the session, T1 appears to provide the clearest example of changes in the domain of practice. In their previous sessions on assessment, as is typical on CELTA courses, the most common topics were international exams (such as IELTS) and testing terminology (such as *washback*). As a result of the LS approach, T1 delivered a session which they reported as being unlike their previous sessions on the topic, as it involved more concrete applications of assessment in the language classroom.

If we then expand the notion of the Domain of Practice from undertaking professional experimentation to an expressed desire to do so, T2 and T3 also described specific plans to replicate this session on their own courses. Of course, it should be noted that planning to do something is not the same as doing it. In addition to delivering the session, T1 also reported that they trialled new ways of using the CELTA assessment criteria, suggesting that LS can encourage further experimentation beyond the research lesson itself:

> On the current course, for example, I have encouraged trainees to share their views on the assessment criteria. (T1)

Domain of Consequence

The third component of the EIMPG, the Domain of Consequence, relates to the outcomes for the trainees. It was beyond the scope of this study to collect direct evidence of changes to trainee LAL. However, we consider it noteworthy that all three trainers mentioned a specific statement from a trainee in which they expressed a wish that the session had been delivered earlier in the course, as the content would have helped her with lesson planning. That this sentiment was particularly salient to the trainers reflects the importance that they place on making input sessions practical, a goal which the trainers felt they achieved. The trainers also observed that it was unusual for such a comment to be made about a session on assessment. Additionally, T1 gathered general end-of-course feedback from all trainees and found that, unprompted, three of the seven trainees wrote positive comments about the session on assessment, one of 40 sessions throughout the course.

External Domain

Finally, the External Domain includes new external influences on trainers. This domain can be considered the one to drive change in the other domains (Perry and Boylan 2018); however as previously stated, the evidence for change in this domain cannot be directly observed. Nevertheless, the

trainers reported benefitting from several different aspects of LS, especially collaboration and peer observation.

Discussion

In the data above, all three trainers reported experiencing some degree of change in a range of areas. With respect to assessment, they all described developing their understanding of assessment principles and their value for trainees. They also gained practical ideas for developing trainees' LAL and saw how these could be used in a session. In general, the LAL development of the trainers was reported to be minor (and was not empirically validated), but extracts from interviews with T2 and T3 suggest that this is due to the nature of the CELTA as an entry-level qualification: 'Pretty much content wise, I knew what the important things would be for the CELTA' (T2). This belief indicates that the LS approach may have even more value when it is explicitly targeted at topics that trainers are less familiar with, on other more advanced courses, or for delivering input that they have not attempted before. These possibilities are alluded to by T3 who saw the value of LS 'as an occasional thing to do as an experiment or to play with a new idea or for something like assessment, where you feel like maybe your session is not meeting [teacher] needs'.

It appears that trainers' External Domains contained a range of influences that led to change in the other domains. The collaborative aspect was frequently highlighted as might be expected, but the observation stage was a rare opportunity that all the trainers valued highly, and the independent research stage led to some further reading that the participant may not otherwise have found. In their study of secondary school teachers using LS, Coenders and Verhoef (2019) similarly found that participants reported benefitting from all aspects of the approach.

Conclusions

Taken together, the findings indicate that all stages of the LS approach are potentially valuable to support the development of trainer LAL in different ways, ultimately leading to growth in trainee LAL. The research supports the findings of Koh (2011), who suggested that AL was most effectively improved through an extended developmental process, and DeLuca, Chavez and Cao (2012) who recommended four activity types found in LS: perspective-building conversations, praxis activities, modelling, and critical reflection.

More generally, the findings suggest that LS can be effective as a tool for trainer development, particularly when trainers have a specific developmental need or aim. In addition, this study demonstrated that LS can be conducted remotely, and that the common practice of having all participants deliver the

same lesson to their own students may not always be necessary. With regard to LAL, this study found that both trainers and trainees can benefit from considering their own experiences of assessment and from viewing assessment as an ongoing and overarching classroom process, rather than the narrower definition of formal testing.

Acknowledgements

We are extremely appreciative of the three trainers who gave their time and effort to participate in this study: Gui Henriques, Patrick Huang, and Sandy Millin. Without their dedication to this process and their insightful contributions, this project would not have been possible.

References

Borg, S (2003) Teacher cognition in language teaching: A review of research on what language teachers think, know, believe, and do, *Language Teaching* 36 (2), 81–109.

Cajkler, W and Wood, P (2016) Adapting 'lesson study' to investigate classroom pedagogy in initial teacher education: what student-teachers think, *Cambridge Journal of Education* 46 (1), 1–18.

Cambridge English (2021a) *CELTA brochure*, available online: www.cambridgeenglish.org/my/Images/107888-celta-brochure.pdf

Cambridge Assessment English (2021b) *CELTA Syllabus and assessment guidelines*, available online: www.cambridgeenglish.org/images/21816-celta-syllabus.pdf

Clarke, D and Hollingsworth, H (2002) Elaborating a model of teacher professional growth, *Teaching and Teacher Education* 18 (8), 947–967.

Coenders, F and Terlouw, C (2015) A model for in-service teacher learning in the context of an innovation, *Journal of Science Teacher Education* 26 (5), 451–470.

Coenders, F and Verhoef, N (2018) Lesson Study: professional development (PD) for beginning and experienced teachers, *Professional Development in Education* 45 (2), 217–230.

DeLuca, C, Chavez, T and Cao, C (2012) Establishing a foundation for valid teacher judgment on student learning: the role of pre-service assessment education, *Assessment in Education: Principles, Policy and Practice* 20 (1), 107–126.

Deygers, B and Malone, M (2019) Language assessment literacy in university admission policies, or the dialogue that isn't, *Language Testing* 36 (3), 347–368.

Fujii, T (2016) Designing and adapting tasks in lesson planning: a critical process of Lesson Study, *ZDM Mathematics Education* 48, 411–423.

Godfrey, D, Seleznyov, S, Anders, J, Wollaston, N and Barrera-Pedemonte, F (2019) A developmental evaluation approach to lesson study: exploring the impact of lesson study in London schools, *Professional development in Education* 25 (2), 235–340.

Koh, K H (2011) Improving teachers' assessment literacy through professional development, *Teaching Education* 22 (3), 255–276.

Nami, F, Marandi, S S and Sotoudehnama, E (2016) CALL teacher professional growth through Lesson Study practice: an investigation into EFL teachers' perceptions, *Computer Assisted Language Learning* 29 (4), 658–682.

Perry, E and Boylan, M (2018) Developing the developers: supporting and researching the learning of professional development facilitators, *Professional Development in Education* 44 (2), 254–271.

Stabler-Havener, M L (2018) Defining, conceptualizing, problematizing, and assessing language teacher assessment literacy, *Studies in Applied Linguistics and TESOL* 18 (1), 1–22.

Villa Larenas, S (2020) *An investigation of the language assessment literacy of teacher educators in Chile: knowledge, practices, learning, beliefs, and context*, PhD thesis.

Weideman, A (2019) Assessment literacy and the good language teacher: four principles and their applications, *Journal for Language Teaching* 53 (1), 103–121.

13 Stakeholder involvement in language examination design: Learning by doing

Dina Vîlcu
Babeş-Bolyai University, Romania

Koen Van Gorp
Michigan State University, USA

Marta García
Universidad de Salamanca, Spain

This case study exemplifies how the reference document *Guidelines for the Development of Language for Specific Purposes Tests* can be used for developing the language assessment literacy/competency of various stakeholder groups in two different contexts of use. The case study focuses on:

- A categorisation of language assessment literacy/competency for different stakeholders
- A presentation of the main characteristics of the *Guidelines*
- A description of how the *Guidelines* have contributed to developing language assessment literacy/competency among teachers and test-takers during the redesign process of (1) an examination of Language for Specific Academic Purposes (LSAP); and (2) an examination for employment and career progression, and for entering undergraduate and postgraduate studies

Acest capitol exemplifică modul în care un document de referință, intitulat *Ghid pentru producerea testelor de limbi pentru scopuri specifice*, poate fi utilizat, în două contexte diferite, pentru dezvoltarea cunoștințelor de bază/competenței în evaluarea limbilor, pentru diferite categorii de persoane interesate. Capitolul include:

- O categorizare a cunoștințelor de bază/competenței de evaluare lingvistică pentru diferite categorii de persoane interesate

- O prezentare a principalelor caracteristici ale *Ghidului pentru producerea testelor de limbi pentru scopuri specifice*
- O descriere a modului în care *Ghidul* a contribuit la dezvoltarea cunoștințelor de bază/competenței în evaluarea limbilor, în rândul profesorilor și al candidaților, cu ocazia procesului de revizuire (1) a unui examen de limbă pentru scopuri specifice academice și (2) a unui examen pentru piața muncii și progres în carieră, ca și pentru admiterea la studii universitare și postuniversitare

Introduction

Over the past two decades, various resources have been created to meet language teachers' language assessment needs, which in turn have enhanced the professionalisation of language education and assessment. Additionally, resources have been created for other groups of stakeholders (test-takers and their families, employers, policy makers), to provide them with information about language assessment. This chapter examines the involvement of two important stakeholder groups, teachers and test-takers, in the redesign of two operational tests in different examination contexts. Both cases illustrate how the use of a test development resource helped to build stakeholders' language assessment competency.

Identifying the LAL needs of different stakeholder groups

Rather than focusing solely on language teachers, language testing researchers conceptualise LAL 'for differing categories of stakeholder needs and levels of proficiency' (Taylor 2013:411). More recently, Baker (2021) proposed 'unpacking' LAL to reflect the various stakeholders' profiles of competence, including different conceptualisations relating to three categories of stakeholder. First, for students and test-takers, Baker suggested the term *(experiential) learning*. She advocated for these stakeholders to actively engage in activities designed to help them become informed test-takers. Second, Baker proposed referring to *expertise and competency* rather than literacy when talking about a variety of professional groups (i.e. teachers, test developers, assessment researchers, admission officers, and policy makers). Third, Baker suggested reserving the term *literacy* for 'the wider public' (i.e. media, test-takers' parents or colleagues), who can benefit from developing an essential understanding of assessment in order to avoid common misunderstandings (like the idea that tests are infallible). These three profiles of competence are not viewed as completely separate from one another.

Aligning with the two first profiles proposed by Baker, this multiple case study shows how two stakeholder groups, test-takers and teachers, were involved in the redesign of two operational tests:

- a Language for Specific Academic Purposes (LSAP) examination used in the medical field in Romania
- the USAL esPro, an examination used for employment and career progression, and for undergraduate and postgraduate education, in Spain.

It describes how the *Guidelines for the Development of Language for Specific Purposes Tests* (edited by the first two authors and described below) inspired us to involve these two stakeholder groups and helped create experiential learning opportunities for the test-takers, and expertise and competency building opportunities for the professional stakeholders.

Foundational documentation in LAL development

Two decades ago, foundational assessment documents like the *Common European Framework of Reference for Languages* (CEFR, Council of Europe 2001) and the Association of Language Testers in Europe (ALTE) 'Can Do' statements (2002) explicitly recognised the need to make language certification processes more transparent to stakeholders. More recent documents, like the CEFR Companion Volume (2018), show how different stakeholder groups can be involved in the assessment development process, shifting the position of test-takers and other stakeholders from a passive one (i.e. receiving information) to a more dynamic and active one (i.e. contributing to the assessment process at various stages).

Two other such documents were produced by ALTE, a professional organisation which has been working towards a better collaboration among different groups of stakeholders. The first document is the *Manual for Language Test Development and Examining* (2011), produced by ALTE on behalf of the Language Policy Division, Council of Europe. It describes the test development cycle of (1) developing, (2) assembling, and (3) delivering tests, then (4) marking, grading and reporting of results, and (5) monitoring and review. The Manual is 'a non-prescriptive document which seeks to highlight the main principles and approaches to test development and assessment which the user can refer to when developing tests within their own contexts of use' (Milanovic 2011:8), and has not only guided many ALTE members in their test design and development activity, but also helped other stakeholders in understanding the intricacies of test development.

However, test developers of examinations of Language for Specific Purposes (LSP) felt that the existing *Manual* needed to be supplemented with

details related to the domain-specificity of LSP tests, such as consideration of the role of subject and context experts (see below) that test developers require. A second booklet was therefore developed, titled *Guidelines for the Development of Language for Specific Purposes Tests* (2018). Both documents are freely available on the websites of the Council of Europe and ALTE. We will focus on this second document in the next section before we show how it has guided the redesign of two operational tests.

Addressing specific challenges in the development of Language for Specific Purposes tests

The *Guidelines* were developed by members of ALTE's Language for Specific Purposes Special Interest Group (LSP SIG), with the help of LSP experts, over the course of multiple SIG meetings. This multi-year process led to a consideration of the following six issues when developing an LSP test.

First, the LSP SIG needed to define and classify LSP tests. If we consider language tests as a continuum from general language proficiency to domain-specific (O'Sullivan 2006), we can identify broad versus narrow LSP tests based on their content. The broad type, like Language for Business or Language for Academic Purposes tests, are closer to the general language proficiency end of the continuum. The narrow, like the Canadian English Language Benchmark Assessment for Nurses (CELBAN), focus on test-takers' functioning in a well-defined domain. As narrow LSP tests present specific challenges to test developers, the *Guidelines* focus on their development, and, more specifically, on the roles of subject, context and language assessment experts. However, broader LSP tests were also considered and referred to throughout the *Guidelines* where appropriate (2018:4).

Second, to help distinguish among the many steps involved in the test development process, the *Guidelines* make a distinction between content and organisational matters in constructing an LSP test. One example of a content issue is the importance of a needs analysis and identification of core tasks, functions, lexis and grammar in the test development phase. An example of an organisational issue is recruiting subject or context experts for the rating of test performances.

Third, throughout the *Guidelines*, we highlight the importance of involving all the different stakeholders, but focus specifically on the involvement of language assessment experts, subject experts and context experts. Subject experts are persons with specialised expertise in a particular field (e.g., a doctor, an air traffic controller); context experts are persons with key expertise in the language of a specific field (like teachers of medical

language) or the needs of the clients or institutions (like health regulators, policy makers). Finding a balance between content and language demands and the insights of all these experts is a challenge for test developers. The *Guidelines* provide clear role descriptions for all experts involved, offering examples (e.g., tests for nurses and police) and practical suggestions (e.g., advantages and challenges of different forms of cooperation between experts) that help test developers make decisions on how to shape the involvement of the different stakeholders and use the advice these stakeholders can bring throughout the test development cycle.

Fourth, for every test, but more importantly for an LSP test, the importance of a strong needs analysis (using different methods and sources – see Long (Ed) 2005) cannot be underestimated. Fifth, a needs analysis also informs issues of authenticity and practicality. Sixth, in the training of item writers and raters, special attention needs to be paid to the role of the subject and context experts. All these challenges are addressed in the *Guidelines*, and each section ends with several key questions that guide test developers in designing an effective LSP test.

The following sections will focus on how the *Guidelines* have been applied in practice to support the redesign of two LSP examinations, and by doing so contributed to the LAL of multiple stakeholders involved in the redesign process. First, we will focus on the redesign of a test of LSAP from the perspective of test producers (teachers) and test-takers (students).

Examination 1. A Language for Specific Academic Purposes test for the biological and biomedical sciences

Background to the project

The *Guidelines* were used to redesign an examination that forms part of the preparatory year in Romanian at the Faculty of Letters, Babeș-Bolyai University, Cluj-Napoca. The academic programme for the preparatory year was redesigned between 2018 and 2020, so this was a welcome opportunity to develop an appropriate assessment tool for the LSAP course of medical language. This is an examination administered to students who intend to pursue medical studies (general medicine, dentistry, pharmacy, biochemistry, veterinary medicine, biology) in Romanian, English or French. It is a high-stakes examination as students are not allowed to graduate if they do not pass it in one of two available sessions. Sixty to 100 students take this examination annually.

For detailed information about the exam redesign process, we refer to the *Illustration for The Guidelines on Language for Specific Purposes*

Tests (ALTE, no date), which is freely accessible on ALTE's website, accompanied by numerous supporting materials (e.g., all research instruments and findings).

Developing the assessment competency of LSAP teachers

Being involved in the redesign process allowed the LSAP teachers, seven in total, to improve their assessment competency. In order to produce the end-of-course examination, they needed to acquire expertise and competency in the assessment of language for medical academic studies. This need was determined in two successive steps. Initially, it resulted from an inter-institutional survey with 33 LSAP teachers from seven universities (Vîlcu and Van Gorp 2018). The responses revealed that all the teachers graduated in the domain of philology (ethnology, literature and general language), with no possibility to specialise in teaching and assessing LSP/LSAP.

The exam redesign process started in ALTE's LSP SIG Meeting in Cluj-Napoca, where the participants discussed the main problems that needed to be addressed through needs analysis and identified the stakeholder groups to be targeted (see Appendix 1).

The two experts who led the redesign process also conducted the needs analysis and designed questionnaires for the LSAP teachers[1], for students, for medical staff (doctors and nurses) and for patients. The responses the LSAP teachers gave to the questionnaires made possible the second step in identifying their language assessment needs. The teachers identified the following main challenges in assessing LSAP: (1) difficulty in deciding how to distribute the points for content vs. language[2]; (2) teachers' insufficient knowledge of medical terms and their proper use; and (3) the selection of adequate texts and creation of good speaking/writing tasks, representative of the target language use (TLU) domain.

Consequently, when assessing LSAP the teachers reported having problems with balancing language and content both in constructing and in marking and rating items and tasks, and deciding on the level of task authenticity in relation to the TLU. As the responses to the LSAP teachers' questionnaires were the first to be collected and analysed, this gave the experts the opportunity to address these problems as part of the redesign process.

1 The seven teachers were part of the team which prepared the general language examination for Romanian language (Levels A1, A2, B1 and B2) for an audit by ALTE. Involvement in this audit process helped them strengthen their competency for general language examinations. However, they felt that their assessment skills in LSP were still not strong enough.

2 Their answers revealed that the distributed points across content and language varied from 70% to content and 30% to language to 20% to content and 80% to language.

The redesign process continued as a constant dialogue between the LSAP teachers and the experts in the LSP SIG: all the steps to be taken (e.g., the interpretation of the students' responses from their needs analysis survey) and problems to be addressed (e.g., authenticity of assessment tasks, the language–content balance) were first discussed with the LSAP teachers; then opinions and recommendations were collected from the LSP SIG members; next, updated findings were presented and discussed during workshops with the LSAP teachers.

Recommendations from the LSP SIG members for a new scoring model providing alternative weights for the four skills and the vocabulary component in the examination were discussed with the LSAP teachers. These discussions also led to a new set of test specifications, reflecting a much more action-oriented approach for the examination. For example, at the level of test construction, role-playing was used for the speaking part (conversations with colleagues and with teachers) and the selected input materials proved to be more adequate (e.g., for reading, articles on medical topics, and for listening, recordings of authentic dialogues: student–student, student–mentor). The students' responses in the needs analysis (see below) proved to be a crucial instrument in the ongoing dialogue between the LSAP teachers and the LSP SIG members. The teachers reported building their assessment competency according to the needs they expressed at the start of the process.

Providing test-takers with an opportunity for experiential learning

Test design starts with a clear understanding of the test-taking population (Brunfaut 2014, *Guidelines*, 2018:8, O'Sullivan 2011). For an analysis of the students' needs, the future test-takers (the students enrolled in the LSAP course at the time that the needs analysis was performed), and the students already enrolled in the University of Medicine, who had taken the old LSAP or similar examinations, were targeted. The purpose of the needs analysis was to collect data about language use in the students' class contexts (courses, laboratories, exams), professional practice contexts (clinics, hospitals), and broader academic contexts (interaction with fellow students, general communication with clinical academics) (Table 1).

Throughout the stepwise collection of data, students had the opportunity to learn about the examinations. The participants in the semi-structured interviews (Step 1) were introduced to the test development stages and were informed about how their responses would help the future exam tasks to be more representative of the TLU domain. Second, they had the opportunity to judge the old examination (in terms of content and level of difficulty) in relation to the real context of their medical studies.

Table 1 Steps in students' needs analysis

Steps	Activities
1.	Semi-structured interviews with 10 students currently in the University of Medicine, who had graduated from the preparatory year
2.	Elaboration of draft questionnaires for students in both the LSAP course and in the University of Medicine
3.	Piloting of the questionnaires with four students from both categories
4.	Sending out the questionnaires
5.	Collecting students' responses (n = 70)
6.	Analysing students' responses with quantitative and qualitative methods

Figure 1 Questionnaire for students in the University of Medicine (fragment)

C. Experience with the language

How important are/were the following activities during your medical studies?	Very important	Important	Quite important	Somewhat important	Unimportant
writing ...					
- notes from written sources	☐	☐	☐	☐	☐
- notes from courses and lectures	☐	☐	☐	☐	☐
- projects on medicine	☐	☐	☐	☐	☐
- essays on medical topics	☐	☐	☐	☐	☐
- research papers/articles on medical topics	☐	☐	☐	☐	☐
- medical documents (e.g. reports on medical cases)	☐	☐	☐	☐	☐
- other writing activities which you consider important (please specify)	☐				

The second group of students who had experiential learning opportunities was the group who helped us pilot the questionnaires (Step 3). This group of students was told that the students already enrolled in the University of Medicine would be asked to rate the importance of some test and teaching activities for their studies (e.g., talking to teachers and fellow students, listening and taking notes in courses, writing specific documents), while the ones currently in the LSAP course would be asked to rate the perceived relevance of these activities for their future medical studies. They were also given examples of the way in which these activities would be reflected in the examination tasks (see Figure 1 for an example for writing activities).

The third group of students were the 70 respondents of the questionnaires (Step 5). Before taking the questionnaires, the experts introduced the students to the relevance of their participation in the redesign process.

To sum up, in redesigning the LSAP examination we provided the teachers with opportunities to build LSP assessment competency through their collaboration with the experts in ALTE's LSP SIG, and for the students with experiential learning opportunities through their involvement in a stepwise needs analysis process.

Examination 2. LanguageCert USAL esPro

In 2017, the University of Salamanca decided to take over the BULATS Business Spanish exam together with its technological partner PeopleCert, renaming it LanguageCert USAL esPro (no date, henceforth esPro) in order to have a Spanish exam for professional purposes. The BULATS (Business Language Testing Service) test was created by a consortium of Cambridge University, Goethe Institut, Alliance Française and the University of Salamanca, and focuses on language use in the field of business. It was redesigned into the esPro exam that, in addition to the business field, also covers professional fields like Spanish for the tourism industry and English for medical professionals. Furthermore, the esPro test covers general Spanish for professional purposes in the workplace as well as educational contexts. It is internationally recognised for employment and career progression, and as entry credentials for undergraduate and postgraduate studies. The test results have important consequences for the candidates, depending on how decision-makers interpret the results.

The University of Salamanca established a clear need for an LSP exam based on a needs analysis which involved different stakeholder groups (CEOs, administrators and teachers) and gauged the TLU domain (e.g., task types, topics, themes and social contexts) and language variation as well as the economic and academic viability of such a test. It also helped differentiate that LSP exam from the two other general Spanish exams developed at the University: DELE for the Cervantes Institute, and SIELE in collaboration with the National University of Mexico, the University of Buenos Aires and the Cervantes Institute.

In this case study, we will focus on one part of the esPro exam: the computer-based adaptive LC USAL esPro Listening & Reading test that assesses listening (18 items), reading (30/32 items) and Spanish language in the intended context, covering all levels of Spanish language proficiency (CEFR A1–C2). Central to redesign of BULATS into esPro were the *Manual* and the *Guidelines*. Additionally, the model proposed by Bachman and Palmer (1996) was followed in order to provide a framework for the concept of test usefulness, that is, a test that can assess language skills in a work environment.

Teachers/Item writers: Building expertise and competency

One of the demonstrable examples of the application of the *Guidelines* was the redefinition of the test and its construct. Based on the needs analysis described above, the TLU was oriented towards the worlds of tourism and higher learning. As a broad examination for professional purposes, experts in the didactics and the teaching of professional subjects in Spanish as a foreign language, as well as experts in the assessment of general and specific-purpose tests, were involved.

Before the redesign of the test started, it was necessary for all teachers who taught Business, Economics, Tourism or Medical Spanish but had no previous experience in assessment to hold professional development sessions twice a week for a month to help them understand key assessment concepts (Step 1 in Table 2).

This phase of professional competency development was foundational so that the parties involved in the development of the exam could work harmoniously. Concepts such as construct, pretesting, specifications or assembling became part of the everyday life of the LSP teachers.

The second part of the training which led to building assessment competency took place during the examination redesign process. Item writers and teachers who checked the editing material had to familiarise themselves with the vocabulary related to the different phases of exam and task construction in order to understand the whole test development process (Step 2 in Table 2). In this process, worksheets were used which contained questions based on those found in various sections of the *Guidelines* (Appendix 2). A total of five expert teachers participated and discussed the checklists to be prepared in advance with the assessment experts, ensuring smooth communication between the two parties. In the following phases of training the teachers familiarised themselves with the tasks included in the examination and with the assessment instruments and procedures (Steps 3 and 4 in Table 2). These last three phases of training especially helped to validate the authenticity of the tasks, as teachers could see whether the students in their classes were competent in dealing with the real-word, professional language use situations in the test

Table 2 LSP teacher training timeline

Steps	Activities
1.	Familiarisation with the assessment concepts
2.	Work with checklists (Appendix 2)
3.	Familiarisation with tasks included in the test
4.	Familiarisation with assessment instruments and procedures

tasks. Therefore, the opportunity was provided for teachers to acquire competence in language assessment as well as competence in specific purposes assessment.

Consequently, we can state that working on assessment competency development of the content experts and teachers was very useful in redesigning the test, but also in making them aware of what the assessment means and of the common processes involved in language certification and classroom assessment. This coordinated effort allowed for feedback that made it possible to move forward in redesigning the exam.

Test-takers: Meeting specific needs

To redesign the test, it was necessary to study the target candidates. As mentioned above, the target group was formed from candidates who were or would be involved in the professional world. Many of them were studying at universities or business schools. In an ongoing effort to ensure the validity and reliability of the exam, information is being gathered from all test-takers. In a post-exam survey, each test-taker indicated to which field they belong and their purposes for taking the esPro. These results allow us to conclude that our candidates take this exam to complete their training and to certify their language level, mostly in the field of economics, business or tourism. This survey continues to be conducted to further verify that the test is adequate for the target audience.

Before the pretesting phase of the examination, training sessions were organised for potential test-takers. We briefed the students about esPro and explained how their contribution would help create a reliable exam. These sessions presented a very positive learning experience because the students worked with tasks that could form part of a future exam, and because the sessions allowed them to do tasks that implied a level of mastery of Spanish in their field of study. As many students did the pretesting sessions voluntarily, knowing that they were not going to receive any official information in return, we understand that they were motivated to be part of the process of creating an exam. They understood that in order to develop a reliable test, test developers need input from (future) test-takers.

Once the test was redesigned, briefing sessions of about 20 minutes were offered in the specific-purpose courses given at the International Courses department of the University of Salamanca. During these briefing sessions, explanations were provided on the exam content and why it is a good exam. In the presentations, concepts such as reliability, score reporting, CEFR scales, 'Can Do' statements or tasks were introduced. To sum up, by creating briefing and professional development sessions for students and teachers,

esPro provided these categories of stakeholder with opportunities to learn about language assessment and build their assessment competency.

Conclusion

We hope that this chapter can contribute to the current general debate on LAL as it is being refined and adapted to different categories of stakeholders, and addresses different profiles of competence. Through the redesigning of two LSP examinations, we have shown how resources for exam development can be used to address the needs of different stakeholder groups, and the problems and issues that might appear in the process of redesigning an examination. Both cases detail how quality test development results from involving developers, teachers, content experts and test-takers, while simultaneously providing affordances for learning. By giving teachers and test-takers an active role in the redesign process, we not only give them a voice, but we also build their language assessment competency and, thereby, empower them. LAL continues to be of great interest to the fields of language assessment and language education alike. By documenting the involvement of different stakeholders in concrete test redesign projects, we add an interpretative layer to the existing resources like the *Manual* or the *Guidelines*. We also contribute to the professionalisation of assessment and build bridges between educational and professional fields.

References

ALTE (no date) *Illustration for The Guidelines on Language for Specific Purposes Tests*, available online: www.alte.org/resources/Documents/Guidelines_Illustration_final.pdf

ALTE (2002) *ALTE Can Do Project*, available online: www.alte.org/resources/Documents/CanDo%20Booklet%20text%20Nov%202002.pdf

ALTE (2011) *Manual for Language Test Development and Examining For Use with the CEFR*, Strasbourg: Council of Europe, available online: www.alte.org/resources/Documents/ManualLanguageTest-Alte2011_EN.pdf

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

Baker, B (2021) *Unpacking the concept of language assessment literacy for all key stakeholders*, plenary talk, New Directions LATAM Online Conference, British Council, July 2021.

Brunfaut, T (2014) Language for Specific Purposes: Current and future issues, *Language Assessment Quarterly* 11 (2), 216–225.

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

Council of Europe (2018) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume*, available

online: rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

LanguageCert USAL esPro (no date) *La última generación de los certificados de español professional*, available online : cursosinternacionales.usal.es/USALesPro/es/

Long, M H (Ed) (2005) *Second Language Needs Analysis*, Cambridge: Cambridge University Press.

Milanovic, M (2011) Introduction, in ALTE (Publisher) *Manual for Language Test Development and Examining For Use with the CEFR*, Strasbourg: Council of Europe, 6–9, available online: www.alte.org/resources/Documents/ManualLanguageTest-Alte2011_EN.pdf

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

O'Sullivan, B (2011) Introduction, in O'Sullivan, B (Ed) *Language Testing: Theories and Practices*, Basingstoke: Palgrave Macmillan, 1–12.

Taylor, L (2013) Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections, *Language Testing* 30 (3), 403–412.

Van Gorp, K and Vîlcu, D (2018) (Eds) *Guidelines for the Development of Language for Specific Purposes Tests. A Supplement to the Manual for Language Test Development and Examining. Produced by ALTE*, available online: www.alte.org/resources/Documents/6093%20LSP%20Supplement%20-%20WEB.pdf

Vîlcu, D and Van Gorp, K (2018) *Developing Resources for LSP Tests: A Reflection*, plenary presentation at ALTE 51st Conference Day, Babeş-Bolyai University, Cluj-Napoca, 13 April 2018.

Appendix 1

Worksheet in the LSP SIG meeting (questions to be answered as a result of the LSAP needs analysis survey and identifying stakeholder groups)

Questions which need to be answered as a result of needs analysis	Groups of stakeholders to be addressed
1. Are there official regulations or laws which define the language test or test delivery conditions in any way?	Government/decision makers/ fund holders
2. What is the social dimension of the test?	
3. What is the test's likely impact?	
4. If the certificate is used for access to a profession or to academic studies, at what stage of the learner's journey towards his/her goal is the language test administered?	
5. Which types of tasks best reflect the needs of the workplace? How authentic do the tasks need to be? How similar to the actual communicative tasks in the workplace?	Medical institution practitioners (doctors, nurses) Patients Teachers in the University of Medicine Medical graduates and students Future test-takers LSP teachers
6. How do the language demands interact with the professional subject knowledge and skills in the field?	
7. What general language skills are needed in the TLU domain that the LSP test targets (i.e., a CEFR level in particular skills, in all skills, etc.)? Is there a threshold level? How has this been set?	
8. What specific language skills are needed for functioning in the workplace? Which domains, topics, themes and social contexts are relevant?	Assessment experts in the domain of LSP LSP teachers
9. What are the lexical, functional and grammatical requirements of the language used in the workplace?	
10. What role does jargon, technical language and professional language play?	
11. To what extent does cultural knowledge need to be incorporated into the test? How will this be done?	

Appendix 2

Checklist for teacher training

Specific purposes course teachers	Answers
What specific language skills are needed for functioning in the workplace?	
Which domains, topics, themes and social contexts are relevant?	
What are the lexical, functional and grammatical requirements of the language used in the workplace?	
What types of tasks are necessary to elicit the required language skills?	
Do the chosen task types represent common practices in the profession?	
How comprehensively is the language of the domain defined or described, and how is it reflected in the items?	
What type of analysis is more adequate for the performance data gathered through piloting, trialling and pretesting, especially if there are not enough candidates for these?	

14 'It tends to be a pretty lonely path …': Exploring the role of a new cross-institutional piloting network for EAP practitioners

Rob Playfair
Birkbeck College, University of London, UK

Eddie Cowling
University of York, UK

This case study investigates the role that a newly established piloting network can play in the work of English for Academic Purposes (EAP) practitioners, through the lens of language assessment literacy (LAL). We explain the origin and design of the network and present findings from interviews with some of the first users of the network from across the globe, all of whom joined to support the development of their own language assessment knowledge and practice. The key findings of the case study are:

- Participants were acutely aware of the benefits of pilot testing but for various reasons had not been able to engage in this activity with their own pre-operational tests
- Participants faced a number of barriers to participating in the network
- Suggestions are given for changes to the network to be more inclusive and facilitate greater engagement from both test writers and EAP managers

Introduction

This case takes place within BALEAP, a professional association for English for Academic Purposes (EAP) practitioners, based in the UK (www.baleap.org).

This association is made up of members who generally work in Higher Education institutions teaching adults using English as a second or other language via foundation year, pre-sessional and in-sessional programmes (Jordan 2002). Although notionally a global association, as of 2021 73% of BALEAP members were based in the UK (BALEAP 2021a). The BALEAP Testing, Assessment and Feedback Special Interest Group (TAFSIG, baleaptafsig.weebly.com) is the site of this case study and both authors are committee members of this group. TAFSIG aims to provide a forum for professional development and community building for EAP practitioners, and to this end organises webinars, blogs and discussion groups on a range of topics related to EAP language testing and assessment, in addition to the Piloting Network (PN) discussed later. EAP tests in Higher Education have diverse purposes, from formative in-class assessment to high-stakes summative exit tests which determine whether students can continue to study and live in the UK. High-stakes tests are accountable to UK immigration law stipulating that students must have proof of a minimum B2 level on the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001) in order to be granted a visa to study in the UK (Home Office 2021:5.10–5.12). The Home Office also conducts institutional 'audits' which include scrutinising the validity of EAP tests (Playfair 2020). These tests are developed by EAP practitioners who often have no formal training in assessment and are engaged in other aspects of EAP work – teaching, marking, personal tutoring – alongside their test development responsibilities. To contextualise the development of the PN, this section outlines how piloting is presented in the literature, the specific challenges faced by EAP practitioners, and how these led to the creation of the PN. We then explain how the theory of language assessment literacy (LAL) has informed this study.

Pre-operational testing is a crucial stage of the test design process in which an evidence-informed case is built to 'assure the test developer and test users that decisions made on the basis of performances on tests are defensible and their consequences beneficial' (Kenyon and MacGregor 2012:298). Piloting is one activity within this stage, in which a test can be used with a sample representative of the final test-taking population, sometimes referred to as *field testing* (Fulcher and Davidson (Eds) 2012) or *pilot testing* (Carr 2011, Dimova, Yan and Ginther 2020). Piloting offers insights on the quality of a test that no other activity can do: how representative test-takers respond. For the purposes of this chapter we will use the terms *pilot testing* or *piloting* as arguably the most frequently used and understandable. It is also for this reason the TAFSIG initiative decided on the name 'Piloting Network', although the longer-term implications for using this label are discussed later.

Relatively little of the literature on test development has considered the development of local tests, which are defined in contrast to large-scale testing

as being embedded in the local context and reliant on local expertise, such as the staff and students within an institution where the test is to be used (Dimova et al 2020). In this chapter we characterise EAP testing as a type of local test in which piloting can take different forms. For example, the test developer may administer the pilot test or give it to the 'host' institution (i.e. the partnering university whose students will take the test) to administer and return to the developers to mark, perhaps to ascertain the quality of the examiner instructions. Another consideration is whether the piloting takes place within or outside of class time or even integrated into the host institution's curriculum, as described in Cowling, Hall, Playfair and Hardman (2021).

Finding students of a similar profile for piloting is rarely considered in the literature, which generally emphasises the benefits of the activity without offering practical guidance for putting it into practice (e.g. Dimova et al 2020, Fulcher and Davidson (Eds) 2012). When such challenges are considered, it is usually in a defeatist tone, with authors conceding that in some cases piloting may just be too difficult (Hughes 2003). For the BALEAP community, meaningful and secure cross-institutional piloting – and all the benefits that come with it – is only possible for those who are able to draw on informal professional networks (e.g. Cowling et al 2021). This exclusivity is detrimental to the development of the EAP field, offering those with connections and experience more opportunity to benefit from piloting than those without.

The concept of the TAFSIG Piloting Network (PN) emerged in one of the group's many informal networking events. The idea was developed further via the BALEAP mailing list discussion threads, with contributors noting the challenge of trialling tests before operational use. The PN was envisioned as an inclusive network connecting interested EAP test developers to pilot tests. To avoid administrative burden, we designed the PN as a way to 'matchmake' suitable institutions, with users of the network providing details about their teaching context, the tests they intend to pilot, and the test-taker population they would like to access and would be able to offer. This information and their contact details are added to a database available to all registered users, who are then encouraged to contact those on the list who meet their needs. At the time of writing there are 48 registered users. Such an innovation has previously been called for in the EAP literature, though conceptualised in top-down terms as a collaboration between EAP practitioners and testing specialists (Schmitt and Hamp-Lyons 2015) rather than among EAP practitioners. The former arguably operates under a different ethos, with EAP practitioners as 'apprentices' to the testing specialists. In contrast, the PN aimed to facilitate contextually grounded (Dimova et al 2020) peer to peer collaboration. Such reciprocal cooperation, we hoped, would contribute to a sense of being joined on the 'lonely path' of EAP test development mentioned by some participants in this study.

Because of the importance of local expertise in EAP test development we feel LAL is a useful lens to identify how the PN could grow and become more responsive to the varying needs of its users. We use Fulcher's (2012) definition of LAL: the knowledge, skills and attributes related to using tests, alongside a familiarity with the range of processes and principles of language assessment. Of particular relevance to this case study is the concept of distinct assessment literacy profiles for different stakeholders, e.g. an EAP test developer compared to a university admissions officer, first suggested by Taylor (2013) and developed by Kremmel and Harding (2020), among others. Associations such as TAFSIG can consider LAL profiles when planning events and designing resources, and tailor initiatives for specific stakeholder profiles. In the small amount of literature on the LAL of EAP practitioners, an awareness of the principles of good assessment practice appears to be widespread (Manning 2013) yet such awareness is accompanied with low confidence in actual test development – including pilot testing – as well as having little opportunity to engage in this area due to the multiple and concurrent roles held in addition to test development, e.g., classroom teacher, coordinator, materials writer (Huang 2018, Manning 2013).

This study aims to inform the future development of the PN by addressing the research question: Why did EAP practitioners choose to join the TAFSIG Piloting Network? In doing so, we hope to shed light on the role of pilot testing in EAP, implementation challenges, and how the PN can activate EAP practitioners' LAL.

Nine EAP practitioners took part in this study, all of whom had signed up to the PN: five based in UK universities, two working at universities in Europe, one in South America and one in central Asia. We have grouped participants by their work responsibilities:

- Three 'EAP tutors' – referred to in this chapter as T1, T2 and T3 with teaching and test development responsibilities.
- Two 'Co-ordinators' – C1 and C2 – developing tests alongside other course management responsibilities such as materials development or teacher training.
- Three 'Managers' – M1, M2 and M3 – overseeing test development, allocating resources, and taking part in some test development but with broader management responsibilities.
- One 'Researcher' participant – R1 – currently undertaking postgraduate study, and who has extensive experience as an EAP teacher and test developer.

The data presented in this chapter is based on individual online video interviews with each of the nine participants, lasting around an hour. Questions focused on participants' knowledge and experience of EAP

test development and their reasons for joining the PN. Using Thematic Analysis (Braun and Clarke 2006) we took an inductive, data-led approach to generating themes from the interviews: reading our notes, reviewing the recordings and discussing our interpretations together. Once we had agreed on themes we returned to the literature to enrich our understanding of our data and 'shed empirical light' on current theory (Yin 2018:38).

Findings

Three themes were generated in response to our research question: Why did EAP practitioners choose to join the TAFSIG Piloting Network? The first two themes illustrate the various motivations participants shared for joining the network, while the third theme explores reasons why these motivations may be frustrated or thwarted.

Recognition of the value of pre-operational testing

Participants cited many benefits of piloting, often going beyond those noted in the literature, including a focus on both test-oriented and more personal benefits.

Theoretical value

Participants agreed that piloting (alongside other pre-testing activities) enhances validity and reliability and can function as a meaningful quality assurance process. Often, this involves analysing the effectiveness of questions and/or difficulty of test items. For most participants we spoke to, however, piloting went beyond item improvement:

> These tests are very high stakes, so [piloting] will help us validate it before it is administered for the first time. But it's not just the information on the item performance, it can also help us spot bias … if any students are disadvantaged … checking instructions and making sure they're not ambiguous. It's that kind of feedback that you might not be able to get from teachers. (C2)

Understanding how we, as educators, might not always 'see the unexpected', and acknowledging the value of going beyond internal staff moderation processes to involving representative learners, was echoed by others. As R1 observed, 'you can carefully craft a test, and it all looks wonderful, but then for some reason it doesn't work, and you'd no way of knowing how or why'.

For some, a perceived lack of external scrutiny in their department's test development practices was a driver to join the PN. Participants spoke

of external examining (EE) processes as seemingly a tickbox exercise, and showed surprise at EE endorsement for what they saw as poor-quality tests: 'They are externally moderated, so someone's coming in and looking at the tests, and saying they're fine, so, I don't get it! [laughs]' (C2). For C2, the light-touch scrutiny of tests they were now responsible for had prompted them to seek to pilot the tests, in order to get what some participants referred to as 'another layer of validation' – such validation they feel they might not be getting elsewhere.

Local and individual benefits

Aside from the contributions piloting makes to the quality of tests, more individualised benefits are also apparent. Participants cited how longer-term piloting can mean a more efficient test-development cycle for their department, as fewer tests need developing once teams are confident they have strong 'tried and tested' assessments, rather than investing time and resources in developing a test which they might need to discard after one operational use. Less experienced test writers spoke of how piloting helped reassure them and instil confidence that they are delivering a robust assessment, even if they were the only one in their department encouraged by such pre-testing measures. There was also a sense of moral obligation, in that the final test-takers shouldn't become what a number of participants called the 'guinea pigs' for high-stakes tests.

The drivers or motivations to carry out test piloting, described above, touch on just some of the reasons that emerged for participants joining the PN. The perceived value of pilot testing meant the potential access to representative test-takers was a major attraction, as participants talked of both the challenges in finding the numbers and appropriate level of students willing to 'try out the test'. Test security was another major pull factor, as piloting within one's own institution was frequently perceived as high-risk, though as we show later, inter-institutional piloting is by no means a panacea as far as test security is concerned. Finally, as explored in more detail later, network participants are signing up for far more than just test piloting with students. Professional development, collegial support, peer-reviewing of tests, and broader collaboration and networking are all added incentives to engage.

Contextual language assessment literacy

Under this theme we grouped comments which addressed the LAL of two key stakeholders. Firstly, participants referred to their own knowledge and skills in test development and the training and preparation they had undertaken before and after taking on their roles – what we term *individual LAL*. Secondly, the perceived collective assessment literacy of colleagues and

managers in their department featured prominently as a driving force for participants to join the network, referred to as *departmental LAL*.

Individual LAL

Many participants in both teaching and managerial roles expressed a sense of not knowing much about assessment when they started their roles. For C2, realising they had responsibilities for test development 'came as a bit of a bombshell' and led to considerable stress and confusion: 'So I did ask the question, what are we doing to make sure these tests are reliable and valid? Like, how can you do that?'

Similarly, in their first testing-related role, T1 found the new test development responsibilities and additional load of learning about assessment overwhelming:

> I wasn't in a good position. I had just been the marker for a number of years, and now this had been dumped on me, I was so new to it, and was trying to read all this stuff [testing literature] at the same time. I was trying to give my rationale but didn't have the experience to back it up.

C2 was relatively new to the institution while T1 was an experienced member of staff in a new role, yet both felt unprepared for testing responsibilities. However, both reported an awareness of their own LAL needs, and sought out ways of meeting them through discussion with colleagues or personal reading. It would seem that for these participants, joining the PN was seen as one of several ways to meet an individual professional need.

Departmental LAL

The quality of the relationships between teachers and managers had a significant effect on participants' motivations for joining the PN. In some cases, the network was seen as an augmentation of well-established departmental assessment practices. Participants in these contexts described regular meetings, peer review, and extensive opportunities for discussing feedback and making revisions. One manager-participant, M1, described such a system:

> … we do tests on potential exam texts, we do Text Inspector [an online text analysis tool: textinspector.com/] … there's a lot of piloting, a lot of checking, standardisation of marking descriptors … everyone hates it to begin with but now it has taken off.

For M1, the PN was a way to offer their staff additional developmental opportunities, further building departmental LAL. Similarly, a teacher, T3, cited an informal arrangement of sharing draft assessments and teaching materials with a trusted manager who provides comments and suggestions.

Unexpectedly for us, these participants in situations where departmental LAL was well developed tended to be interested in using the PN as a means of getting new ideas and meeting colleagues with a shared interest, rather than meeting any immediate need to pilot.

Participants working in contexts where departmental LAL appears less developed saw the potential of the network to fill this gap, as a way to assuage their feelings of frustration or isolation. In many cases the teacher-participants attributed this to managers' lack of experience in developing assessments and the impact on work allocation. C2 felt their department took the quality of the tests for granted and that questioning their validity wasn't welcomed. R1 noted a similar lack of interest when describing their previous EAP centre's attitude towards piloting: 'The perception [from managers] is "why would you wanna do that? It's a waste of time, money, resources … It's just a test".' Contrasting a previous manager with their new one, T1 noted the effects of managerial LAL on the team:

> [New Manager] understands the process. Previous managers have been too focused on the product and had little understanding of what it takes to create a valid, reliable product; I was constantly pressured with "how quickly can you get this test done …", whereas [New Manager] is more "how realistic is this time scale?" I realise I'm talking about [New Manager] a lot but I really feel that the type of manager has an immense impact on this role and the value assigned to it.

Barriers to piloting

Despite the recognition of the value of piloting, participants spoke of specific psychological and practical barriers which had prevented them from engaging in pre-testing activities. These barriers were constructed from threats, fears and assumptions about piloting. An exploration of these can point the way to making our PN more effective and inclusive.

Security fears

Sharing tests across institutions creates more opportunities for them to be released into the public domain. In some cases, participants shared experiences of previous unsuccessful inter-institutional piloting attempts with security issues resulting in the team being unable to use the test after piloting. C1 felt that the most secure way to pilot would be in face-to-face settings and if done online, this would require extensive liaison with the hosting institution to ensure that appropriate measures were put in place such as proctoring or lockdown browsers. This relates to the next sub-theme.

Intra- and inter-institutional relationships

Pressure from within their own institution was a barrier for T1. When asked why they had not engaged in piloting before, they explained, 'we've had, you know, the pressures from the international office, of wanting "new tests, new tests, new tests", and, erm, so we just feel we're scrambling to keep up with that'.

A number of participants mentioned the need for a close and honest relationship between both institutions in order to pilot effectively. C1 referred to this as establishing 'ground rules' in relation to the what, where, when, who and how. This could include whether test-takers participating in the pilot receive feedback on their performance, from whom, in what form and by when. Other institutional responsibilities such as test administration and communication with test-takers were cited as areas that would need to be agreed upon. This prospect of establishing such a relationship posed a barrier for some as it involves a time commitment before any piloting can take place in what for many is an already overburdened workload.

The fear of being exposed to criticism was mentioned by a few participants as a potential barrier when discussing sharing a test to pilot with peers in other institutions. M1 hinted at this when they mentioned the potential additional effort needed for such discussions: 'If you looked at our [marking] rubrics you'd probably think that the weighting for each band is strange, but you have to explain the context.' As mentioned earlier, M1's motivation for joining the network was mainly for informal sharing rather than detailed piloting and we suggest that this mention of a 'need to explain yourself' could be a barrier to greater participation in piloting. For a more direct illustration of this theme, during their interview C2 shared a test they were working on, prefacing it by saying 'please don't judge me!' We feel these displays of vulnerability give an insight into how test developers could be encouraged into participation in the network, perhaps through lower-stakes activities such as discussion groups or online forums. This will be discussed further in the following section.

Implications and discussion

The three themes generated from the interviews provide insight into the interlinked reasons that participants joined the network and possible barriers to its success. Based on these, we now offer suggestions for how the PN can develop further as well as avenues for further research into LAL in EAP contexts.

Raise awareness of the principles *and* processes of pre-operational testing

All participants recognised the value of pilot testing in line with previous studies (e.g., Manning 2013). This is unsurprising, considering the self-selecting sample drawn from users of the PN. As we have seen, an individual's awareness of this is not sufficient to enable piloting to happen and some participants felt a need for more guidance on the processes (Fulcher 2012) of carrying this out. Also, the role of managers in EAP testing appears to be an important, and at times constraining, factor. Due to their complex and wide-ranging role, we assume most managers have strong contextual knowledge – observed by Dimova et al (2020) as a benefit of local test development – yet some still seem to be unable to create the conditions for adequate local test development. We feel this deserves some exploration. In all but one of the participants' contexts, assessment roles were held by a minority of staff and in most cases as one part of a number of other non-assessment-related duties, a state which we believe is typical within EAP centres. In other words, most EAP teachers are not involved in developing assessment. As such, it is perhaps unsurprising that some EAP practitioners are able to gain management positions without such knowledge or experience. However, it does raise the question of what kind of LAL is relevant for those in such roles who aren't involved directly in test development, but oversee and manage those who are. This is a nuance not yet addressed in the LAL literature, which includes the role of 'language assessment/test developer' but not those who oversee and influence this development (Kremmel and Harding 2020). The closest to this we are aware of is the 'policy maker', but this role is further from the day-to-day test development process than the EAP manager – for example, policy makers have been conceptualised as 'non-practitioners' such as government officials or doctors (Pill and Harding 2013). Considering the influence EAP managerial roles have on tests and those who develop them, we call for further exploration of the power dynamics and knowledge bases of managers in local test development and how such knowledge differences between managers and test developers can be reconciled.

A practical implication for us, suggested by participants, is to produce a resource aimed specifically at EAP managers which contains an evidence-informed account of timescales and resources needed to produce a quality language test, and emphasises the benefits that the participants in this study have noted. Such a document, if supported (or even co-created) by respected professionals in the field and professional organisations such as BALEAP and UK Association for Language Testing and Assessment (UKALTA), would be a valuable tool to support more realistic resource allocation for quality test development. Further engagement by BALEAP through more detailed assessment-related descriptors in individual and institutional accreditation schemes could also encourage more of those in positions of power to develop their own knowledge in this area.

Create opportunities for test-development skills development

Interviewees lamented a lack of affordable and accessible training in the field of language test development, particularly running and analysing pilot tests, and there was clear appetite for upskilling in this area. Similar to other studies (Huang 2018, Manning 2013) a lack of confidence in organising test piloting and handling post-pilot test data prevented some colleagues from exploring piloting opportunities. In response TAFSIG is planning a workshop on collaborative test development and another on item analysis. We also hope to showcase some of the network's 'success stories', to provide inspiration and guidance for those uncertain about piloting.

As per the recommendations of interviewees, we will also compile literature and provide guidance on various steps within the test design process. With increased awareness of the value of pre-testing and the time and resources needed to create good quality tests, it is hoped managers will also utilise and/or refer staff to such resources.

Develop a guide document for PN users

There appears to be an appetite for both formal piloting and more informal collaborative test development such as peer review, discussion groups and consultation. Making these options visible to participants may help diminish some of the barriers noted in the previous section. A 'piloting network guide' could help users to align expectations and reduce anxiety about participation. This could include options along a continuum of formal and informal collaboration supported with example terms of reference for approaches to test security, marking of pilot tests, and provision of feedback. An online test writers' discussion forum could allow users of the network to pose non-piloting related questions and seek answers, advice and ideas from colleagues in the language testing community. In addition, we could change the name of the network, which we felt may have led potential users into believing the network's parameters were more restricted than our intentions. Something more general such as 'test development network' might be more inviting.

Conclusion

This case study set out to inform the development of a new Piloting Network by exploring the motivations of some of the initial users. Participants were positive about the benefits of piloting, yet due to an intersecting range of factors were unable to engage in this activity to their satisfaction. These factors include individual LAL and the LAL of those in their department, managers in particular. Even having joined the network, participants still faced a number of barriers to piloting, such as fears relating to test security, limited time and

peer criticism. We suggest a number of strategies the network can employ to address some of these barriers and also call on greater collaboration with the wider EAP and test development communities. Two suggested areas of future research are to explore the applicability of the PN for other contexts beyond the BALEAP association and to develop a better understanding of all stakeholders' – especially managers' – LAL in EAP test development.

References

BALEAP (2021a) *Chair's report 2020–21*, available online: www.baleap.org/wp-content/uploads/2021/04/BALEAP-Chair-Report-AGM-2021.pdf

BALEAP (2021b) *BALEAP Testing Officer's Report*, available online: www.baleap.org/wp-content/uploads/2021/04/BALEAP-Testing-Officer-Report-AGM-2021.pdf

Braun, V and Clarke, V (2006) Using thematic analysis in psychology, *Qualitative Research in Psychology* 3 (2), 77–101.

Carr, N (2011) *Designing and Analyzing Language Tests*, Oxford: Oxford University Press.

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

Cowling, E, Hall, M, Playfair, R and Hardman, W (2021) *An Inter-institutional TAFSIG Piloting Network – a case study and discussion*, available online: baleaptafsig.weebly.com/archived-events/an-inter-institutional-tafsig-piloting-network-a-case-study-and-discussion

Dimova, S, Yan, X and Ginther, A (2020) *Local Language Testing: Design, Implementation, and Development*, Oxford: Routledge.

Fulcher, G (2012) Assessment literacy for the language classroom, *Language Assessment Quarterly* 9 (2), 113–132.

Fulcher, G and Davidson, F (Eds) (2012) *The Routledge Handbook of Language Testing*, New York: Routledge.

Home Office (2021) *Student Sponsor Guidance Document 2: Sponsorship Duties*, available online: assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1022004/Student_Sponsor_Guidance_-_Doc_2_-_Sponsorship_Duties_2021-09-29_FINAL.pdf

Huang, L S (2018) A call for critical dialogue: EAP assessment from the practitioner's perspective in Canada, *Journal of English for Academic Purposes* 35, 70–84.

Hughes, A (2003) *Testing for Language Teachers* (Second edition), Cambridge: Cambridge University Press.

Jordan, R R (2002) The growth of EAP in Britain, *Journal of English for Academic Purposes* 1 (1), 69–78.

Kenyon, D and MacGregor, D (2012) Pre-operational testing, in Fulcher, G and Davidson, F (Eds) *The Routledge Handbook of Language Testing*, New York: Routledge, 295–306.

Kremmel, B and Harding, L (2020) Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the language assessment literacy survey, *Language Assessment Quarterly* 17 (1), 100–120.

Manning, A (2013) *EAP Teacher Assessment Literacy*, unpublished PhD thesis, University of Leicester.
Pill, J and Harding, L (2013) Defining the language assessment literacy gap: Evidence from a parliamentary inquiry, *Language Testing* 30 (3), 381–402.
Playfair, R (2020) In-house EAP Testing and the UKVI – 5 case studies, *TAFSIG Blog*, available online: baleaptafsig.weebly.com/blog/in-house-eap-testing-und-the-ukvi-4-case-studies
Schmitt, D and Hamp-Lyons, L (2015) The need for EAP teacher knowledge in assessment, *Journal of English for Academic Purposes* 18, 3–8.
Taylor, L (2013) Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections, *Language Testing* 30 (3), 403–412.
Yin, R K (2018) *Case Study Research and Applications: Design and Methods* (Sixth edition), Thousand Oaks: Sage Publications.