

Language Assessment Literacy
and Competence Volume 1:
Research and Reflections from
the Field

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 Khabbazzbashi

Assessing Academic Listening: The Case of IELTS

John Field

Language Assessment Literacy and Competence Volume 1: Research and Reflections from the Field

Edited by

Beverly Baker

University of Ottawa, Canada

and

Lynda Taylor

CRELLA, University of Bedfordshire



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

© Cambridge University Press & Assessment 2024

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

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 978-1-009-80231-4

Cambridge University Press & Assessment has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Contents

Acknowledgements	vii
Series Editors' note	ix
Notes on contributors	xiii
1 Reflecting on an apprenticeship journey in language assessment literacy <i>Lynda Taylor</i>	1
Section 1	
Empirical Research Studies	
2 Developing learning-focused feedback practices: An evaluation of English for Academic Purposes teacher feedback literacy at a Sino-foreign university <i>Joseph Arthur Davies</i>	25
3 Discerning the language assessment literacy of EFL teachers in Uzbekistan: A social and sociohistorical teacher cognition inquiry <i>David L. Chiesa</i>	46
4 How do teacher educators learn about and teach language assessment? <i>Salomé Villa Larenas and Tineke Brunfaut</i>	66
5 Supporting K-12 educators' language assessment literacy via resources informed by validation frameworks <i>Ahyoung Alicia Kim, Shireen Baghestani, David MacGregor and Pauline Ho</i>	86
6 Enhancing teachers' and test-takers' assessment literacy? Insights from test preparation for the digital TestDaF <i>Sonja Zimmermann, Leska Schwarz, Anja Peters and Günther Depner</i>	110
7 An online self-study course for Celpe-Bras raters and examiners: Induction into assessment practice and reflection <i>Margarete Schlatter, Matilde Virginia Ricardi Scaramucci and Juliana Roquele Schoffen</i>	129

- 8 Language assessment literacy in a workplace environment – an exploratory study from the testing of language proficiency in aviation 148
Neil Bullock
- 9 Shaping a language testing curriculum: Insights from an oral history of a Master’s programme 168
John Pill, John Bandman, Raffaella Bottini, Tineke Brunfaut, Natalya Davidson, Geisa Dávila Pérez, Luke Harding, Yejin Yung, Santi B. Lestari, Camilo Ramos Gálvez and Olena Rossi

Section 2

Scholarly Reflections

- 10 Sign language assessment: Common misconceptions and the need for language assessment literacy development 195
Franz Holzknrecht and Tobias Haug
- 11 A conceptual framework to contextualise local language assessment literacy 204
Newton Paulo Monteiro
- 12 Practical but not atheoretical: A reflection on the promise of concept-based instruction in teacher assessment literacy development 213
Xuan Minh Ngo
- 13 Collaborative assessment cultures and the development of LAL 223
Henrik Bøhn and Dina Tsagari
- 14 ‘A course of our own’: Reflections following a language assessment literacy (LAL) course 231
Ofra Inbar-Lourie
- 15 Advancing professionalisation: The achievement of language assessment literacy 239
Albert Weideman
- 16 Assessment literacy theory: Pragmatics, complexity and transdisciplinarity 250
Christopher DeLuca

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

Shireen Baghestani is a PhD student at Iowa State University in Ames, Iowa, USA. Her research interests include argument-based test validation, rater cognition, and language assessment literacy of test users (particularly those without formal assessment training). Previously, she worked for several years as an IELTS examiner and item writer for the TOEFL iBT. She also worked in the Middle East, where she taught English as a Second Language/English as a Foreign Language (ESL/EFL) to adults, undergraduates and high school students. In the summer of 2020, she worked as a research intern at WIDA.

John Bandman is a PhD student in Applied Linguistics at Lancaster University. Besides teaching ESL pedagogy at Hunter College, New York, he teaches full-time at Bergen Community College in the New York City area. He also teaches student success, business, and culinary and beverage management – his other careers.

Henrik Bøhn is Associate Professor of English didactics at the University of South-Eastern Norway, where he works as a teacher, supervisor and researcher in language education. He has earlier worked as an English teacher in secondary school in Norway and has been involved in curriculum development and fostering teacher professionalism in the Norwegian context. His research interests include language assessment, assessment literacy, learning strategies and intercultural learning.

Raffaella Bottini is a Lecturer in Corpus Linguistics in the Department of Linguistics and English Language at Lancaster University. She specialises in the application of corpus methods to the analysis of L2 English to inform language testing and language teaching. Her research interests include corpus design, corpus grammar, and register analysis.

Tineke Brunfaut is Professor in Linguistics and English Language at Lancaster University, where she specialises in language testing. Her research has been published in journals such as *Language Assessment Quarterly*, *Language Testing*, *Assessing Writing* and *TESOL Quarterly*. She regularly conducts language test development, training and consultancy work for professional and educational bodies.

Neil Bullock is currently Vice-President and General Secretary for the International Civil Aviation English Association, and is based in Switzerland. He has been an English teacher for many years, working mainly in professional and specific-purpose domains, and has been involved in test development programmes in the world of aviation and general-purpose language. He has published and presented regularly on teaching and testing matters on language proficiency in aviation. His research interests include broadening the scope of communicative competence, assessment literacy and technology-mediated assessment.

David L. Chiesa is Clinical Assistant Professor of Language and Literacy Education at the University of Georgia, USA. He worked for many years on various language assessment and testing projects with the US Department of State's Office of English Language Programs in conjunction with government organisations in the Republic of Uzbekistan, Mainland China, and Mongolia. His research interests include language assessment literacy, second language teacher education, and the integration of technology with language teaching, learning and assessment.

Natalya Davidson is a PhD student in the Department of Linguistics and English Language at Lancaster University. Her doctoral study investigates the construct of multiple-text comprehension in integrated reading-to-speak test tasks. She also works as an English teacher at Zefat Academic College, Israel.

Joseph Arthur Davies is Senior Lecturer of English Language and Assistant Director for graduate English for Academic Purposes (EAP) at Duke Kunshan University's Language and Culture Center (LCC), People's Republic of China. He has managed and taught across various English language programmes within Sino-foreign higher education contexts. He currently leads a graduate EAP team to develop and teach interdisciplinary EAP courses, and teaches an undergraduate language teaching methodology course. Joseph's research interests include student and teacher feedback literacy, classroom language assessment, and EAP pedagogy.

Geisa Dávila Pérez is a doctoral student at Lancaster University. Her research focuses on identifying and understanding features of English as a Lingua Franca (ELF) in the context of speaking tests, particularly in the Graded Examinations of Spoken English administered by Trinity College London. Findings may help gain insights into the variation of ELF features across proficiency levels and its implications for language assessment. Her research interests include speaking assessment, multilingual assessment, and examiner/rater training.

Christopher DeLuca (PhD, Queen's University) is an Associate Dean at the School of Graduate Studies and Postdoctoral Affairs and Professor in Educational Assessment at the Faculty of Education, Queen's University, Canada. He leads the Classroom Assessment Research Team and is Director of the Queen's University Assessment and Evaluation Group. His research examines the intersection of assessment, curriculum, and pedagogy from sociocultural frameworks. His work largely focuses on supporting teachers in negotiating these critical areas of practice to enhance student learning for all.

Günther Depner studied Educational Science, English, Language Teaching/Learning Science, and German as a Foreign Language in Bochum, Germany. At g.a.s.t. (*Gesellschaft für akademische Studienvorbereitung und Testentwicklung e.V.* [Society for Academic Study Preparation and Test Development]) he is involved in test development and trainings on topics related to this area. His research and training interests are interlanguage, formative assessment, developing reading and writing competencies. He holds a PhD in developing argumentative writing skills from the Ruhr University Bochum, Germany.

Luke Harding is Professor in Linguistics and English Language at Lancaster University. His research interests are in language assessment and applied linguistics more broadly. His work has been published in a range of academic journals. He is the current Vice-President of the International Language Testing Association (ILTA).

Tobias Haug is Professor in Sign Language and Participation of the Deaf and Hearing-Impaired at the University of Teacher Education in Special Needs in Zurich, Switzerland. His research focuses mostly on issues related to sign language assessment and sign language interpreting. He has (co-)authored many articles and book chapters on sign language assessment and co-edited the book *Handbook of Language Assessment Across Modalities*, published by Oxford University Press in 2022.

Pauline Ho is a PhD candidate in the Department of Educational Psychology at the University of Wisconsin-Madison. She works as a Project Assistant for WIDA on several projects focusing on the reclassification and identification of English language learners.

Franz Holzknacht is Advanced Researcher at the University of Teacher Education in Special Needs in Zurich, Switzerland. He holds an MA and a PhD in Language Testing from Lancaster University and an MA in teaching English and Sports from the University of Innsbruck. He has worked in

language test development and research for many years and has published numerous journal articles and book chapters. His research interests are mainly in language assessment, particularly in sign language and L2 listening assessment.

Ofra Inbar-Lourie lectures in the multilingual education department in the School of Education at Tel Aviv University. She has published widely on language assessment, particularly with regard to language teachers and language policy, and in her writings coined and promoted understanding of the ‘Language Assessment Literacy’ concept. She was also active in establishing the LAL SIG in ILTA. Her current interest is in the merge between assessment literacy and the language of instruction in academia.

Yejin Jung is a PhD researcher in the Department of Linguistics and English Language at Lancaster University. Her research interests are the assessment and testing of speaking, L2 pragmatics, interactional competence, and corpus-assisted investigations of test discourse. Her PhD project examines the linguistic features that occur when L2 speakers express their opinions in interactive speech and identifies the relationship between the features and proficiency level in L2.

Ahyoung Alicia Kim is a researcher at WIDA, University of Wisconsin-Madison, where she examines the language development of K-12 multilingual learners. Her research interests include language assessment, child bilingualism, second language literacy development, and computer-assisted language learning. Alicia received her doctorate in Applied Linguistics from Teachers College, Columbia University. Prior to joining WIDA, she was a post-doctoral fellow in child bilingualism at Cornell University and a language assessment researcher at Seoul National University. She has also taught TESOL and English as a Second Language (ESL) courses in the US and abroad.

Santi B. Lestari is a doctoral student in Linguistics at Lancaster University. Her doctoral study investigates ways of operationalising the reading-into-writing construct in analytic rating scales and their impact on raters’ rating. Her research interests include L2 writing assessment, integrated tasks and rater cognition.

David MacGregor, PhD, is an Assessment Researcher for WIDA. Prior to coming to WIDA, he worked on various language assessment projects at the Center for Applied Linguistics (CAL). Much of his work at CAL was focused on WIDA-related projects. He also has over a decade of experience teaching

ESOL in Turkey, Spain, and the US. He is interested in issues related to test validity, standard setting and computer-adaptive testing.

Xuan Minh Ngo is an Associate Lecturer in International Education and TESOL at the University of St Andrews, UK. In his recently completed PhD project at the University of Queensland, Australia, Minh examined the development of language assessment literacy among novice English teachers through the lens of Vygotsky's sociocultural theory. His research interests lie in the intersection of language assessment, language policy, and teacher education.

Newton Paulo Monteiro is a lecturer of English, Portuguese and Linguistics at Alfredo Nasser University Centre (UNIFAN), Brazil, where he leads an in-house committee for local assessment and evaluation issues. He has been in charge of conducting assessment projects for language and general education as well as assessment training sessions for teacher education. His research interests include local language test development, language assessment literacy in context, and management of assessment projects.

Anja Peters studied English, German and American Studies in Frankfurt, Southampton and Minneapolis. After her doctorate, she worked as a lecturer at several universities in the UK. Since March 2018, she has been working at g.a.s.t. (*Gesellschaft für akademische Studienvorbereitung und Testentwicklung e.V.* [Society for Academic Study Preparation and Test Development]) as a test validation officer. She is responsible for the listening and reading sections of the paper-based Test of German as a Foreign Language (*Test Deutsch als Fremdsprache* – TestDaF), develops test preparation materials and conducts training on topics related to test development.

John Pill is a Lecturer in Language Testing in the Department of Linguistics and English Language at Lancaster University. He is currently director of studies for the MA programme in Language Testing. His research interests include specific-purpose language testing, the scope and definition of language constructs in particular contexts, and teaching and communicating about language assessment.

Camilo Ramos Gálvez is a doctoral student in Language Testing at Lancaster University, where he currently investigates the diagnosis of academic reading in English as a foreign language. His research interests are in language assessment and English language teacher education. He works as a lecturer in Santiago, Chile.

Olena Rossi has an MA and a PhD in language testing from Lancaster University. Her research interests are in test design, item writing, and assessment literacy for test stakeholders. Her doctoral research study looked into an online induction item writing training course to gain insights into item writing skills and their development.

Matilde Virginia Ricardi Scaramucci is a full professor in the Department of Applied Linguistics, *Universidade Estadual de Campinas (Unicamp)*, Brazil, and a volunteer researcher in the Postgraduate Program of Applied Linguistics of the same university after her retirement in 2018. She is one of the developers of the Certificate of Proficiency in Portuguese for Foreigners (Celpe-Bras) and member of the exam scientific committee since 1994. Her research interests include teacher education, language assessment literacy, language assessment of integrated tasks and washback.

Margarete Schlatter is a professor in Language Teacher Education at the *Universidade Federal do Rio Grande do Sul (UFRGS)*, Brazil. She is one of the developers of the Certificate of Proficiency in Portuguese for Foreigners (Celpe-Bras) and has worked for many years in the development of online self-study courses for learning Portuguese as an additional language, and for teacher development in teaching and assessing reading, writing and speaking. Her research interests include teacher education, language teaching and assessment practices, and materials design.

Juliana Roquele Schoffen is an associate professor and a permanent researcher at the *Universidade Federal do Rio Grande do Sul (UFRGS)*, Brazil. She has been a member of the Certificate of Proficiency in Portuguese for Foreigners (Celpe-Bras) scientific committee since 2008. Her main research interests include language assessment and teacher education.

Leska Schwarz studied Instructional Design and German Linguistics in Freiburg, Germany. She is a test validation officer at g.a.s.t. (*Gesellschaft für akademische Studienvorbereitung und Testentwicklung e.V.* [Society for Academic Study Preparation and Test Development]), where she is involved in task development and scoring as well as trainings on topics related to test development and the digital Test of German as a Foreign Language (*Test Deutsch als Fremdsprache – TestDaF*). Her research interests include the testing of listening comprehension, and the possibilities of automated or technology-assisted scoring of short-answers.

Dina Tsagari is Professor at the Department of Primary and Secondary Teacher Education, Oslo Metropolitan University, Norway. Her research interests include language testing and assessment, materials design and

evaluation, differentiated instruction, multilingualism, distance education, learning difficulties and inclusive education. She is the editor and author of numerous books, journal papers, book chapters, project reports etc. and coordinates research groups and EU-funded and other research projects.

Salomé Villa Larenas is Academic Researcher at Universidad Alberto Hurtado, Chile, where she conducts research on language test impact and language assessment literacy. She is co-founder and immediate past president of the Latin American Association for Language Testing and Assessment (LAALTA). Her research interests include language assessment literacy development in the teaching profession and language testing in teacher education.

Albert Weideman is Professor of Applied Language Studies and Research Fellow at the University of the Free State, South Africa. He has recently published *Assessing Academic Literacy in a Multilingual Context: Transition and Transformation* (2021, Multilingual Matters). He focuses on language assessment design and developing a theory of applied linguistics.

Sonja Zimmermann is the Head of Product Development, with a focus on testing, at g.a.s.t. (*Gesellschaft für akademische Studienvorbereitung und Testentwicklung e.V.* [Society for Academic Study Preparation and Test Development]). She has many years of experience in language assessment and validation projects, mainly for productive skills. Before joining g.a.s.t., Sonja worked as a lecturer for the German Academic Exchange Service in the Russian Federation. She holds a PhD on integrated writing assessment from the University of Bremen, Germany.

1 Reflecting on an apprenticeship journey in language assessment literacy

Lynda Taylor

Centre for Research in English Language Learning and Assessment (CRELLA), Bedfordshire

Introduction

Some of the roots of this edited volume stretch back more than three decades – to the late 1980s. That was the time when I first began working as a member of the test research and development team within the English as a Foreign Language (EFL) Division at the University of Cambridge Local Examinations Syndicate (UCLES) in the UK.

By that point in my professional career as an English language teacher, I had become increasingly aware of the growing role played by testing and assessment in the life of my students, including the potential impact of test scores on their future educational opportunities and life chances. I had spent a decade involved in routine placement testing and classroom assessment within the language school institutions where I taught, in both London and Cambridge, as well as in preparing learners to take external examinations at the end of their English language courses. These examinations included Cambridge's First Certificate in English and Certificate of Proficiency in English tests, the listening/speaking assessments offered at that time by ARELS (the Association of Registered English Language Schools), and the relatively new and innovative English Language Testing Service (ELTS), first introduced in 1980 and precursor to the International English Language Testing System (IELTS) released in 1989. (Detailed information on all these language tests can be found in a number of other volumes published in the Studies in Language Testing series, including Hawkey 2004, Davies 2008 and Weir, Vidaković and Galaczi 2013).

Towards the end of the 1980s, as I began to work more closely with Cambridge, first as a freelance writing examiner and item writer, and later as a test developer and researcher, it became clear to me that my undergraduate modern languages studies followed by postgraduate teacher training to become an English language teacher during the 1970s had not included much that was related to the theory and practice of language testing and assessment. A sound basis in linguistics and language pedagogy was clearly

useful, but this was not really sufficient for tackling the complex challenges of language test design, development, delivery and validation. I quickly realised that I lacked some of the knowledge, skills and understanding of core principles and practice that were needed to develop and implement good quality assessment tools for language learners – tests that would serve both individuals and society well with regard to future educational and employment opportunities.

At that time, I could not have given a name to this sense of professional inadequacy and personal anxiety, and I had little idea of how to address or resolve it. Looking back now, of course, I realise that I was aware of what our field later came to conceptualise as a need for ‘language assessment literacy’, later abbreviated to LAL. This sense of inadequacy and anxiety in the face of test design and development demands in the early 1990s was the start of an ongoing personal commitment to learning more about language assessment in order to try and ‘do language testing well’, as both a competent practitioner and a responsible researcher. The experience also caused me to reflect on how such a process of learning, competence-building and professional development takes place, not just for practitioner stakeholders like myself in those early days, but for other assessment stakeholders (e.g. teachers, learners, curriculum developers, university admissions officers), many of whom can find themselves directly (or indirectly) involved in language testing and assessment, sometimes with significant professional responsibility but with limited access to personal professional development opportunities in this area.

In this introductory chapter I aim to capture and describe aspects of my own personal narrative and journey into LAL in the hope it may reveal and highlight some important and relevant themes in the development of language assessment literacy and competence. As such, the chapter sets the scene for what follows in the edited collections of research papers, case studies and reflections contained in the two volumes which Beverly and I have had the pleasure and privilege of assembling and bringing to the field. This chapter will explain the rationale for the two edited volumes and their overall content and structure. First I will review the concept of LAL that has been developing over the past two decades, noting how the process of learning about language assessment and developing the associated expertise takes place within different assessment stakeholder constituencies, often using different approaches and methods. I shall consider the many different stakeholders in language assessment, with their respective and often varying needs for assessment-related knowledge and skills, and reflect on how such knowledge and skills are, or can be, acquired. Towards the end of this chapter, I will briefly introduce the eight empirical research studies and seven scholarly reflections which make up this first of the two volumes, discussing themes that emerge from these

contributions to highlight issues that might benefit from further attention and be potential avenues for future research.

Recognition of multiple stakeholders in language testing and assessment

The 1990s saw growing interest among language testers in specific issues of test washback and impact, as well as in language testing ethics and social responsibility more broadly (see, for example, Alderson and Wall 1993, Alderson and Hamp-Lyons 1996, Hamp-Lyons 1997). This was especially true in relation to the high-stakes English language tests being widely used around the world at the time, e.g., IELTS and the Test of English as a Foreign Language (TOEFL). Interest in the ethics of language testing prompted growing awareness of the multiple ‘stakeholders’ in language testing and the ways in which different types of stakeholder can be involved in an assessment enterprise, sometimes with unique perspectives. Such involvement may be either direct or indirect, whether in the complex process of test design, development and delivery, or in the critical areas of test preparation or test score interpretation and use. The list of stakeholders can be surprisingly long. It includes not only test developers and researchers, but also teachers, learners, parents and care-givers, coursebook writers, curriculum designers, teacher trainers and educators, educational policy makers, ministry officials and others.

In 1997, Pauline Rea-Dickins published one of the first peer-reviewed articles to explore this dimension by addressing the question ‘So, why do we need relationships with stakeholders in language testing?’, published in the journal *Language Testing*. Rea-Dickins (1997) identified five particular stakeholder categories: learners, teachers, parents, government and official bodies, and the marketplace. Interestingly, she did not include item writers or test examiners, or even academic researchers, in her stakeholder list. Nevertheless, her paper offered a valuable starting point for considering the washback and impact of language tests on an extended network (or web) of individuals and groups, all of whom have a professional interest (or stake) in the language testing enterprise and who may therefore need to develop some measure of knowledge and skills to support their involvement.

In a contribution to Cambridge’s new *Research Notes* publication in August 2000, I made a preliminary attempt to identify a wider range of test stakeholder types which included the item writer and test examiner communities (Taylor 2000). Building on the earlier work of Rea-Dickins, my aim was to examine more closely the nature of the relationship that exists between a test provider, such as UCLES, and the many groups in language education and wider society who have a stake or share an interest in the use and value of any language test the examination board might be

offering in the public domain. I explored the complex patterns of stakeholder relationships within language testing, highlighting what I perceived to be differing models used to describe these relationships. The prevailing model at that time tended to separate stakeholders into ‘producer’ and ‘consumer’ categories, a sort of binary ‘test-maker versus test-taker’ model. A more extreme perspective proposed a sociopolitical view of the language testing enterprise in which one party exercises power over another, sometimes expressed in a sort of ‘perpetrator–victim’ model. A fine-grained analysis, however, revealed a far more complex community of participants and set of relationships than can be represented by a simplistic two-sided model.

By expanding Rea-Dickins’ original categorisation, it was possible to conceptualise a broader range of stakeholders including coursebook writers, publishers, school owners/administrators, test centre staff, examiners, employers and academic researchers in language testing and assessment. The last of these categories was rapidly expanding at that time, both within university linguistics departments and within professional testing organisations (see Taylor and Green 2020 for more discussion of this). I was particularly interested in understanding how a major English language test provider included these stakeholders in the overall assessment process, from the design, development and delivery of its tests, to the interpretation and use of test scores, to the monitoring, quality assurance and revision processes for its assessment products, and even to the provision of special arrangements for candidates unable to take a standard test format due to illness or disability, whether temporary or permanent.

Working during the mid-1990s as an invited language assessment consultant with national test development agencies around the world also highlighted other types of stakeholders who may be concerned with the introduction of a new test into society, e.g., from the domains of politics, philology and pedagogy as well as those with concerns over human rights and social justice, and even the news media. What an individual stakeholder, or constituency of stakeholders, needs to understand about an assessment and its use is likely to vary considerably according to their specific context and role within society.

In my 2000 article on ‘Stakeholders in Language Testing’ I briefly described five specific aspects of stakeholder-relevant activity undertaken by Cambridge EFL in its role as a major test provider: the methodology for test revision; the quality assurance system for oral examiners; the role of a code of practice; support and information for test stakeholders; and the provision of arrangements for test-takers with special needs. (Looking back, I wonder if this was perhaps an early – and somewhat naïve – attempt on my part at defining some sort of LAL construct!) In the same issue of *Research Notes*, my colleague Nick Saville discussed what it means for an examination board to be accountable in its relationship with the many stakeholders associated

with its tests, highlighting the importance of transparency and positive engagement (Saville 2000). Interestingly, considerations of stakeholder needs and expectations, and the tailoring of communication to engage intended groups, have come to the fore once again as highlighted by Chalhoub-Deville and O'Sullivan (2020) in their recent discussion of validity.

Those short discussion papers published in 2000 helped to clarify my own thinking, and served as a springboard for reflecting on and writing about language assessment literacy and competence in the years that followed. Thinking about what different test stakeholders need to understand or learn to develop the competence and confidence to do their work well (and to experience personal and professional satisfaction in that work) leads on to considering how such experience or expertise are best encouraged or facilitated.

Recognising the needs of different stakeholder groups

My own interest in understanding and addressing stakeholder needs had largely focused on the needs of item writers for EFL tests. In the UK context at that time, test item writers were typically experienced EFL teachers or coursebook materials writers who also worked freelance for examination providers, such as Cambridge or Trinity College. In my role as a test researcher with Cambridge in the early 1990s, I was invited to lead and coordinate teams of such externally commissioned item writers to develop test materials for their suite of EFL exams, including new test development projects at the time such as Cambridge Advanced and a revised IELTS.

Recalling my own lack of assessment knowledge and skills when starting out as an item writer just a few years previously, it seemed sensible to organise short one- and two-day item writer training courses. The courses introduced key principles and practice in language testing and assessment, drawing on some of the publications emerging at that time, including Henning (1987), Hughes (1989), Davies (1990), Bachman (1990) and Weir (1990). The primary aim of such courses was to better equip item writers with the knowledge and skills needed to develop good quality test tasks and items, and to revise existing exams or develop new ones. The idea was to strengthen understanding of latest developments in applied linguistics and measurement theory (e.g., Rasch analysis was beginning to be used for item banking), as well as to create a sound practical awareness of the organisational test development cycle (i.e., item drafting, editing, revision, pretesting/trialling, analysis). A further aim was to help item writers understand how and where they fit within the wider frame of reference for a test's development, from its initial conception through to operational delivery. The approach reflected a steady move towards the professionalisation of all aspects of

language assessment. This shift was also reflected in the growing number of more practically oriented books on language testing published during the mid-1990s, including Weir (1993) Alderson, Clapham and Wall (1995) and Bachman and Palmer (1996), followed a few years later by a series of specific skill-focused volumes, e.g., on assessing writing or vocabulary (cf Read 2000, Weigle 2002). Over little more than a decade during the 1990s and 2000s, the body of theoretical and applied literature available to support the development of language assessment literacy grew rapidly, and this growth has continued.

Addressing stakeholder needs through a community of practice approach

Reflecting upon my own experience as an item writer, and as a Chair of various item writing teams over those years, it seemed important to try and encourage among my fellow freelance item writers a strong sense of community which could help to build and maintain our shared item-writing competence. For example, when IELTS was being extensively revised in the mid-1990s, I implemented a localised, course-based training approach to help prepare and support IELTS item writers in both the UK and Australia. The aim was to create competent and confident teams of item writers for the newly revised IELTS modules for Reading, Writing and Listening, supported by a face-to-face item writer training programme. This led to the creation of a comprehensive Handbook for Item Writers containing practical guidelines on every aspect of test item development. A similar professional development programme was created to retrain and support the worldwide community of IELTS Writing examiners in 1995 and the examiner-rater cadre for the revised IELTS Speaking test introduced in 2001. Once again, these initiatives were facilitated through face-to-face and self-access training sessions combined with extensive documentation.

This perspective and approach to addressing stakeholder needs resonated with the concept of ‘community of practice’ proposed in the early 1990s by social scientists and educators Jean Lave and Etienne Wenger. Their seminal volume published in 1991, entitled *Situated Learning: Legitimate Peripheral Participation*, provided a useful theoretical framework for understanding how effective learning can be social (rather than just individual) and is often achieved by participating with others in a ‘community created over time by the sustained pursuit of a shared enterprise’ (Wenger 1998:45). This view also resonates with the earlier work of philosopher Donald Schön (1983), who explored how organisational learning systems can exist in domains beyond the formal educational context, e.g., in the workplace, and encourages an emphasis upon becoming a reflective practitioner.

Lave and Wenger's 1991 situated learning model was premised on a process of mutual engagement within a community of practice, and this had certainly been my own experience during the late 1980s. I had begun as a novice item writer and test developer on the periphery, and gradually acquired the necessary knowledge and skills for the role through a form of corporate and mutual apprenticeship. Over time, as I moved closer to the centre of the community, I became equipped (and confident enough) to take on a team leadership role or assist in the design of a new test. What is noteworthy about this personal experience is how little it depended upon following any sort of taught course or on completing a qualification – approaches that are perhaps more typical of the academy. My own learning about language assessment was largely social and interactional, through participation with others in the hands-on activity of test writing, review, editing and redrafting. Nevertheless, that early community of practice experience did prompt me to undertake additional individual studies in language assessment at both Master's and doctoral level in the early 1990s to consolidate and deepen my expertise.

The literature on communities of practice recognises several key elements that contribute to successful and effective teams in their respective workplace contexts. Such elements include: generating and appropriating a shared repertoire of ideas; developing various resources, such as documents, tools, vocabulary and symbols, all of which help to carry the accumulated knowledge of the community and make it easier to induct new members; and building positive relationships that enable open, honest discussion within the group and a respectful critique of each other's contribution. In practice, although conversations in test review and editing meetings could sometimes be difficult and egos could occasionally feel bruised, a willingness to review and critique the work of others, and in turn have one's own work reviewed and critiqued, was an essential part of the learning process. In fact it was essential for producing good quality test material in advance of trialling and pretesting stages in test development. Once again, this element resonates with Schön's (1983) 'reflective practitioner' model, involving both reflection-in-action and reflection-on-action.

My later interest in promoting 'language assessment literacy' (though I would not have used the term at that time) stems directly from personal and professional experience of working with test item writers, and later writing and oral examiners. I found that a situated learning model within a community of practice can be an effective method to develop the knowledge, skills and principles that make up competence – or literacy – in language assessment, and that was a useful piece of learning for me.

The emergence of language assessment literacy (LAL)

Stiggins (1991, 1995, 2014) is generally credited as being among the first to use the term ‘assessment literacy’ (AL) though the concept related to a broad educational context rather than the specific field of language education. Stiggins used the term to denote the characteristics of an assessor who might be more or less ‘assessment literate’, i.e., capable of differentiating between high- and low-quality assessment instruments and their outcomes, and able to implement appropriate changes if improvement to testing tools or procedures is needed. This included having access to reliable measurement data for decision-making purposes. The term was largely restricted to describing what teachers and instructors (and possibly also policymakers) need to know about assessment matters in the course of their professional work.

During the 2000s some members of the language testing and assessment community adopted the ‘assessment literacy’ term as a useful concept in their own professional domain (Stoynoff and Chapelle 2005, Inbar-Lourie 2008, Malone 2008), especially as the role and influence of language testing increased in areas such as international higher education, professional recruitment (e.g., healthcare), and migration and citizenship. They rightly argued that language teachers, instructors and administrators need some measure of assessment training if they are engaged in selecting, administering, interpreting, and sharing results of large-scale language tests produced by professional testing organisations, or in developing, scoring, interpreting, and improving classroom-based assessments. It was perhaps not surprising that Stiggins’ original phrase was expanded to the term ‘language assessment literacy’, leading to the shorter, and more convenient, acronym: LAL.

Given the growing role for language proficiency assessments as part of social policy in an increasingly globalised world, *The Annual Review of Applied Linguistics* chose to focus their 2009 volume on the topic of language policy and language assessment, guest edited by Professor Bernard Spolsky. The volume included a chapter on developing assessment literacy (Taylor 2009) in which I set out to review efforts to ‘promote understanding of assessment within the field of applied linguistics and education and society more broadly’ (2009:21). I also reflected on how the international language testing community could ‘encourage the sharing of the core knowledge, skill, and understanding that underpin good quality assessment as widely and accessibly for the benefit of all’ (2009:21). That chapter noted a number of emerging trends: the growing numbers involved in language assessment; the increasing professionalisation of language testing and assessment; attempts to conceptualise LAL and its components; and recognition of the need to improve LAL both within and beyond the language testing profession.

In June 2011 the international language testing community's annual conference – Language Testing Research Colloquium (LTRC) – took place in Ann Arbor, Michigan, on the theme of 'Half a Century of Language Testing'. Among other things, the conference celebrated Robert Lado's seminal volume *Language Testing* published 50 years earlier in 1961. The programme included a symposium on the topic of LAL with contributions from acknowledged experts on the subject: Cathie Elder, April Ginther, Glenn Fulcher, Meg Malone and Ofra Inbar-Lourie. My role as symposium discussant was to draw together thematic threads from the papers presented in response to several key questions: *Who are the stakeholders requiring AL/LAL? What sort of content input do they require? In what specific domains/contexts? When is the best time for this to happen? What methods are likely to be most effective?* My concluding remarks to the symposium highlighted the need for more research to be undertaken and published in four main areas: understanding and defining the AL/LAL construct; the nature of the language/discourse we use when engaging with non-specialists; identifying, evaluating and responding to varying user needs; and understanding how AL/LAL grows/matures over time.

As a follow-up to the 2011 symposium, a special issue of the journal *Language Testing* was commissioned containing a set of five papers by Scarino (2013), Malone (2013), Jeong (2013), O'Loughlin (2013) and Pill and Harding (2013). Once again, I was invited to draw together a number of themes in a concluding discussion paper entitled 'Communicating the theory, practice and principles of language testing to test stakeholders: some reflections' (Taylor 2013). My aim in this particular paper was threefold: i) to revisit some of the themes addressed in the 2011 LTRC symposium; ii) to consider how these had been further explored or developed through the special issue papers; and iii) to reflect on some future directions for our thinking and activity in this area.

Since 2013, that paper in *Language Testing* has been widely read and regularly cited in discussions about LAL. This is encouraging and it is also pleasing to feel that one may have contributed something useful to the debate! Some of the ideas I shared seem to have captured researchers' imaginations and inspired a great deal of applied research seeking to explore the likely components of LAL and the extent to which these differ across stakeholder groups (see, for example, Kremmel and Harding 2020). Baker (2019) suggested that representing the idea of differential LAL profiles for differing stakeholders (which I did using spidergrams in Figure 2 – 2013:410) allowed us, among other things, to visualise the complementary nature of each stakeholder's LAL. However, it is also true that parts of the paper – including that particular graphic – have sometimes been misinterpreted or mis-represented, in inappropriate and unhelpful ways.

In an online discussion at the Reading Group of the ILTA LAL Special Interest Group in 2020, I was invited to talk about the 2013 paper and its impact. I reflected on what I believed I had covered in my contribution, namely: acknowledgement of Lado's legacy; recognition of multiple literacies nowadays; the need for urgent empirical research in AL/LAL; linkage of the five special issue papers to the four research strands that came out of the 2011 symposium; differentiation of selected stakeholder groups mentioned in the other papers (test writers, classroom teachers, university administrators, as well as professional language testers); and my personal reflections on likely differentiation in the respective needs of these groups. As I explained to the LAL Reading Group, I did not seek to offer a comprehensive theory-based and empirically grounded framework or model for the development of AL/LAL for specific stakeholder groups; nor did I try to make claims about the usefulness of my particular characterisation for theoretical or practical research purposes. In fact, the oft-critiqued Figure 2 in my paper was deliberately not referred to as either a framework or a model. To underpin this point, it may be worth quoting directly from the actual paper (note I have added my own emphasis here to highlight the speculative nature of what I was trying to communicate):

Figure 2(a-d) *attempts to illustrate* what different assessment literacy *might* look like for these 3 groups and for the community of professional language testing experts ... the labelled dimensions on the eight axes (i.e. knowledge of theory, technical skills, etc.) *are hypothesized* from the discussion of *possible* AL/LAL components across various papers in this special issue, while the values (i.e. 0–4) *are hypothesized* according to the different stages of literacy *suggested* by Pill and Harding. *The diagrams are for illustrative purposes only, to show how it might be possible to conceptualise and represent differential AL/LAL; the actual characterisation is naturally open to debate.* (2013:409–410)

An unfortunate (and unforeseen) consequence of that 2013 paper was that some readers were tempted to remove Figure 2 from its original context and purpose and to overinterpret it as a fully fledged framework or comprehensive model ripe for empirical investigation. Over the years, I have noted with some dismay numerous references in the LAL literature to the 'Taylor model', including criticisms of its (lack of) completeness! I am happy to set the record straight here, and I believe the confusion or overinterpretation associated with that 2013 paper is worth commenting on because it highlights the danger of taking something out of context and using it for a purpose for which it was not designed. After all, as language assessment specialists we are keenly sensitive to the danger of doing that with a test and we often caution others against it!

Perhaps most importantly, we need to be reminded that matters of context, purpose and perspective are critical when considering the LAL requirements of stakeholder communities. I could not have come up with an authentic and comprehensive model of LAL on my own because I can only represent one viewpoint on any language assessment endeavour. The LAL needs and priorities of any specific stakeholder group or population cannot necessarily be anticipated and prescribed by professional language testers alone (even if they are the so-called ‘experts’). For example, the stakeholder category of ‘language teachers’ is not a homogeneous population; the knowledge, skills and understanding that teachers of young learners in the language classroom need may differ considerably from those needed by university-level teachers in higher education contexts. Similarly, the LAL needs and priorities of policymakers or administrators working with language assessments in well-resourced parts of the world may well differ from what is needed (or possible) in countries where access to resources, whether time, money, personnel, etc., is more limited. This can be particularly true in parts of the Global South or in regions that have suffered long-term social and political instability. It makes sense that LAL needs and priorities are best clarified and negotiated with members of a given stakeholder group in a respectful and reciprocal relationship, as the work of Baker and Riches (2018) clearly demonstrates in the context of Haiti.

Rationale for publishing a volume on language assessment literacy and competence

A key rationale for publishing these two complementary volumes on language assessment literacy and competence is that we believe enough conceptual and empirical work has been undertaken over the past decade to enable us to cast our view back and reflect upon where we have arrived as a field and where we might go next. Since publication of the special issue of *Language Testing* in 2013, there has been an explosion of interest in LAL, with a proliferation of research studies, conference presentations, journal articles and, to date, a relatively small number of published volumes on the topic (see, for example, *Language Assessment Literacy: From Theory to Practice*, edited by Dina Tsagari (2020)). Most publications have tended to focus on categorising the knowledge and skills needed and what needs to be acquired by specific groups or constituencies of those who use tests or test scores in their local contexts of work or study.

April 2019 saw the establishment of the Language Assessment Literacy Special Interest Group within ILTA (known as LALSIG) with the following mandate: to expand the constituency of ILTA to include a broader range of assessment stakeholders; to encourage partnerships and information sharing between language assessment specialists and other language assessment

stakeholders – to the mutual benefit and LAL development of all parties; and to promote communication among researchers working in the area of LAL, cultivating the development of international research partnerships. The list of LAL-related publications originally collated by LALSIG members has been taken over by the ILTA Graduate Assembly (GSA) and integrated within the ILTA Bibliography, which is freely available online. Interestingly, the first volume of the ILTA Bibliography (covering 1990–1999) contained no separate LAL-related category. In the second volume (covering 1999–2020) all LAL-related publications were included within the General section of the publication. The third volume, however, covering 2021–2022, includes a dedicated Language Assessment Literacy subcategory under the section entitled Social Aspects of Language Assessment. At the time of writing (August 2023), the LAL section lists a total of 50 references!

Research efforts to date have tended to focus heavily on what sort of competence language teachers might require for assessing their students, reflected in a number of empirical studies investigating teacher assessment literacy in differing contexts around the world. A smaller number of studies have investigated the understanding of language assessment issues needed by admissions tutors in international higher education contexts. In addition to these two important constituencies, however, as we saw earlier in this chapter there are many other stakeholder groups with differing responsibilities and in differing contexts, all of whom require an understanding of language assessment to varying degrees – an understanding that will be strongly shaped by multiple features of the context in which they operate (e.g., education, policymaking, public media). As the field of LAL continues to evolve and mature, Beverly and I believed there was scope for a collection of edited papers that would provide a state-of-the-art publication bringing together knowledge and experience from around the world. Cambridge University Press & Assessment's Studies in Language Testing series offered a good publication venue for such a project, especially as no title thus far in the series has specifically addressed the topic of language assessment literacy and competence.

In July 2021 we issued an open call to solicit a wide range of papers representing different contexts around the world and reflecting different approaches to learning and skills development. We invited a range of contributions: conceptual pieces on the nature of language assessment knowledge and its development; studies of professional competency development and situational learning within the workplace; case studies outlining innovative methods to develop language assessment competence; and shorter vignettes from test-takers and other stakeholders on critical learning moments, sharing their experience in their own words. Accounts of collaborative work among stakeholders were especially welcome, as well as contributions from those whose voices are less often heard in the global

debate. The number of quality submissions received led to a decision to edit and publish two complementary volumes on the theme of language assessment literacy and competence: Volume 1 focusing on research studies and reflections from the field; and Volume 2 profiling case studies from around the world.

The first part of Volume 1 contains a set of eight longer chapters reporting on empirical research studies conducted in different contexts around the world, while the second part gathers together a set of seven shorter pieces in which authors reflect upon LAL issues in their respective contexts. As they share their thoughts and experiences from differing perspectives, their reflections may help to stimulate and push conversations forward in new directions.

Key themes emerging from the research studies and reflections in Volume 1

It is not difficult to discern multiple and recurring themes or ‘threads’ that weave in and out of the research studies and reflections presented in this volume, creating a tapestry from which we can stand back in order to view the bigger picture. In the final part of this introductory chapter I aim to draw out some of the most salient threads to consider what they tell us about how understanding of language assessment literacy and competence continues to evolve with implications for theory and practice in language assessment.

Language assessment literacy and competence are highly context-based in nature

This fundamental principle seems to be axiomatic. Awareness of and sensitivity to context is critical in LAL needs analysis and in shaping LAL development programmes. Some years ago, I suspect we tended to think of LAL, and LAL development, in a somewhat monolithic or homogenous sense, i.e., as a domain comprised of standard content components which apply regardless of context or population and which follow a fixed developmental trajectory. In reality, LAL and its successful development are proving to be highly context-sensitive and context-dependent. There can be no ‘one-size-fits-all’ approach to developing LAL – no prescriptive curriculum or ideal programme as we perhaps once imagined or believed. Many of the contributions in this volume highlight how a sound understanding of local contextual factors – within a domain, an institution, an industry, a society or some other bounded phenomenon – is essential for making decisions about the content or focus of any programme to develop LAL in/for that context. Several authors note the varying LAL requirements and purposes that may need to be acknowledged according to different stakeholder constituencies and changing circumstances across a range of different contexts.

Newton Paulo Monteiro (Chapter 11), for example, highlights the complex professional roles and identities that typically shape LAL development in the Brazilian higher education context. He proposes a conceptual framework to reflect on local professional experiences in developing LAL there, arguing that LAL should be seen as contextualised practices related to local professional roles (PRs), or functional identities. LAL thus develops under the influence of the various drivers and constraints that stakeholders take on or are subjected to in these professional roles. In a different context, **Franz Holzknacht** and **Tobias Haug** (Chapter 10) point to the highly specific LAL needs of the sign language community. The authors address common misconceptions about sign language assessment, identifying needs in LAL development for professionals working in sign language education – a specialised yet under-represented stakeholder group in the literature. They highlight challenges associated with language assessment literacy development in minority linguistic communities and suggest future work to foster language assessment literacy development in sign language communities in particular.

Drawing on data collected as part of a wider oral history project, **John Pill, John Bandman, Raffaella Bottini, Tineke Brunfaut, Natalya Davidson, Geisa Dávilla Pérez, Luke Harding, Yejin Jung, Santi B. Lestari, Camilo Ramos Gálvez** and **Olena Rossi** (Chapter 9) describe how the curriculum of a part-time, online, distance-based Master's programme in language testing at Lancaster University was conceptualised and modified over almost two decades. They discuss its role in developing the knowledge and skills of emerging language testing specialists and the impact of learning on students' subsequent professional practice, noting how the programme topics (as well as the personnel delivering the syllabus) changed over the years. Their account speaks to the dynamic and evolving nature of any knowledge/skills/principles base within a wider historical sociopolitical, educational, epistemological, ethical and moral context or framework. We might note here that issues of equity and social justice, accessibility, decolonisation, sustainability, etc. have recently assumed greater prominence in the field of language testing and assessment in recent years, and it may be topics such as these which shape the LAL curriculum in the years to come.

Stakeholder beliefs and attitudes can play a significant factor in LAL

Another noticeable thread running through several contributions is the importance of understanding and attending to the personal beliefs, attitudes and prior experiences of stakeholders, particularly teachers and learners, since these can significantly impact not only levels of assessment literacy and professional competence, but also choices of how best to approach

development activities. For example, **Joseph Arthur Davies** (Chapter 2) explores the feedback beliefs and practices of Higher Education English for Academic Purposes (EAP) teachers at a Sino-foreign joint venture university, to help develop our understanding of teacher feedback literacy for teachers in contexts where competing cultural values and expectations can impact on students. His research highlights the complex cross-cultural challenges that can emerge for EAP teachers and he uses his research outcomes to inform a teacher feedback literacy framework, discussing its implications for HE EAP teachers wishing to develop their own teacher feedback literacy as part of their professional assessment competence.

In a different regional and sociocultural context, **David L. Chiesa** (Chapter 3) seeks to bridge the two academic fields of language teacher cognition and LAL in his study of the local institutional and larger sociocultural context for language teachers in Uzbekistan, with its complex political history and social traditions. Examining an under-researched population of EFL teachers, he explores knowledge, beliefs, and feelings within the sociocultural contexts for teaching, learning and assessment to discern how the historical and sociopolitical factors underpinning EFL teachers' beliefs and attitudes in Uzbekistan can directly impact on their assessment understanding and practice.

Ofra Inbar-Lourie (Chapter 14) and **Xuan Minh Ngo** (Chapter 12) share engaging, first-person accounts of their own experiences in testing and assessment, one as a teacher educator and the other as a test-taker, highlighting the impact of their personal beliefs, attitudes and experience in shaping their LAL. **Inbar-Lourie** draws on her experience of teaching an LAL course to pre-service teachers of different languages as part of a teacher certification programme in a university in Israel. She highlights the important role of critical language assessment and active relevant learning and research in the acquisition of LAL and language assessment identity, stressing the development of LAL expertise as a situated paradigm from both language and assessment perspectives. **Ngo** reflects on his struggles in delivering an effective language assessment course in the Vietnamese context, seeking to make such courses more practical and accessible without sacrificing their theoretical rigour. He proposes that teacher assessment literacy development can be seen as a process of concept formation, involving the synergy of empirical and scientific concepts. His frank and unflinching take on the struggles we face in supporting teacher development in language assessment competency is both refreshing and engaging.

Stakeholder communities can often be interdependent

In terms of the categorisation of different stakeholder groups, it is encouraging to see greater attention being paid to some groups previously

under-represented in the literature, including learners/students, teacher educators, test raters/examiners and professionals within the aviation industry.

Sonja Zimmermann, Leska Schwarz, Anja Peters and Günther Depner (Chapter 6) evaluate the use of a new approach for test preparation based on the principles of learning-oriented assessment (LOA). In the context of a standardised proficiency test for admissions purposes in Germany, they explore how the use of this approach for teachers and learners can in turn promote the development of LAL, especially through classroom test preparation activities, and they stress the importance of the link between what happens in the test and what happens in the target language use (TLU) domain. **Salomé Villa Larenas and Tineke Brunfaut** (Chapter 4) explore the development of language assessment competence among teacher educators in Chile, helpfully shifting our focus beyond the community of teachers to those who train them. In a study of how teacher educators learn about language assessment themselves and how they teach this to the next generation of language teachers, the authors examine the concept of teacher learning as seen through the lens of the sociocultural theory of learning, highlighting the important role played by modelling and reflective practice.

Margarete Schlatter, Matilde Virginia Ricardi Scaramucci and Juliana Roquele Schoffen (Chapter 7) analyse the design of an online self-study course for the Certificate of Proficiency in Portuguese for Foreigners (Celpe-Bras) raters/examiners, describing how the theoretical framework of the course design is based on induction into assessment practice and reflection. Practice and feedback are integrated to promote the development of knowledge and skills in language assessment, an approach the authors suggest may be relevant for other exams, with implications both for increasing the validity and reliability of a test and for supporting language professionals wishing to develop expertise in additional language teaching, learning and assessment practices.

Neil Bullock (Chapter 8) explores the role of LAL within the international aviation industry's system for assessing the language proficiency of air traffic controllers and pilots. He discusses the importance of the knowledge, skills and experience of diverse stakeholder groups within the aviation industry (e.g., pilots, air traffic controllers, language trainers), noting the extent to which these groups are often interdependent and how their literacy levels may influence the effectiveness of the system. Interestingly, even within this specialised professional domain and constituency, Bullock observed the LAL needs to differ among the various players involved. Based on analysis exploring stakeholder groups' perceptions of the knowledge and skills required for their roles, he offers suggestions on how greater assessment literacy can be encouraged among stakeholders in the aviation industry.

Collaborative engagement within and between stakeholder groups can be key

Studies such as those referred to above suggest greater recognition nowadays of the interdependence of some stakeholder communities, as well as the importance of studying and facilitating language assessment development of different groups in combination – both within and across teams. It is interesting to note the importance of key individuals and strong teams in the effective development of LAL and professional competence, perhaps echoing the communities of practice paradigm I observed in the 1980s as an essential part of my own LAL journey.

In this regard, **Ahyoung Alicia Kim**, **Shireen Baghestani**, **David MacGregor** and **Pauline Ho** (Chapter 5) consider collaboration between test producers and K-12 educators within the US public school system to improve test understanding and score interpretation. They investigate how an argument-based validation framework, specifically the Assessment Use Argument (AUA) by Bachman and Palmer (2010), might be used to inform the development of resources designed to support the LAL of educators serving kindergarten to grade 12 (K-12) English learners (ELs). They explore the accessibility and relevance of such an argument-based validation framework for non-specialist test stakeholders, and the extent to which it may need to be mediated in some way in order to be practically useful for the purposes of wider stakeholder engagement.

Henrik Bøhn and **Dina Tsagari** (Chapter 13) stress the value of promoting a collaborative culture for quality assessment within a different state education system, this time in Norway. They discuss the importance and value of seeking to ensure collaboration in the development of teacher LAL within the Norwegian and similar contexts, among subject teachers, students, school leaders and teacher educators. **Pill et al** (Chapter 9) also note the impact of instrumental individuals and teams (not just courses and programmes) in the effective development of LAL and professional competence, while **Davies** (Chapter 2) discusses the need for shared responsibility between students and teachers in the assessment feedback process. The communities of practice paradigm allows for that important dynamic of people moving inwards and outwards to sustain and refresh the enterprise, not just within a community of practice but across different, though related, communities. Once again, we see an interesting shift away from a strongly compartmentalised view of LAL as residing within professional silos or specific sub-populations, and instead being distributed across a network of stakeholder teams or communities, mutually supporting and learning from one another, as witnessed and reported by Baker and Riches (2018).

Language and discourse need to be audience- and context-appropriate

One of the most interesting points of discussion to emerge in the field of LAL in recent years concerns the use of terminology. There remains some debate about whether the original term ‘assessment literacy’ is actually a better general term to use in language testing, rather than narrowing the concept down to focus on language assessment. Although most authors seem comfortable with the more domain-specific ‘language assessment literacy’, some authors stand out as preferring a broader characterisation of the concept that is detached from any sense of disciplinarity. **Christopher DeLuca** (Chapter 16), for example, challenges the notion of assessment (and thus assessment literacy) as discipline-dependent, and instead advances an argument for assessment as inherently and necessarily transdisciplinary, pragmatic, and complex. While recognising that disciplinary context does shape assessment in practice, he proposes that a broader conceptualisation of assessment is needed beyond a purely disciplinary view.

Interestingly, the widely accepted term ‘language assessment literacy’ has recently been problematised by some as conveying an unhelpful ‘deficit’ connotation when contrasted with its antonym ‘illiteracy’. So it may be that differential terms are preferable in our discourse, at least for certain stakeholder communities? It could be argued, for example, that a better term to use for teachers might be ‘competence’ rather than ‘literacy’. Assessment competence is analogous to other types of professional competence that teachers need to acquire; in other words, assessment knowledge, skills and understanding represents just another type of professional competence. For the general public, however, the term ‘literacy’ might remain suitable (Baker 2021), aligning as it does with contemporary concepts such as ‘digital literacy’ and ‘information literacy’. Taylor and Harding (2020) and Taylor (2023) offer a fuller discussion of the importance of selecting the right sort of language and discourse to raise awareness and improve understanding of assessment matters among the general public and non-specialists. Beynen (2023) proposes a pluralised term, ‘assessment literacies’, to describe the multi-faceted nature of the assessment awareness that students entering higher education in Canada might need, and it will be interesting to see if this pluralised form is taken up more widely in the future.

Concern over the terminology and discourse used in the field of assessment echoes one of the four issues I raised at the LTRC 2011 Symposium on LAL, where I called for more research to be undertaken and published on the nature of the language/discourse we use when engaging with non-specialists. Despite considerable progress over recent years on three of the four issues I highlighted at that event (i.e., understanding and defining the construct of AL/LAL, evaluating and responding to varying stakeholder needs, and

understanding how AL/LAL develops over time), the language we choose and use for communicative engagement with non-specialist stakeholders remains a major area where further thinking and research are needed.

Geographical reach and representation are steadily expanding

Perhaps one of the most encouraging and exciting aspects of the Volume 1 contributions is the extent to which they represent so many different parts of the globe – a feature which is mirrored in the complementary Volume 2 presenting case studies from around the world.

In Volume 1 the sites for research and reflection include Brazil, Chile, China, Germany, Israel, Norway, South Africa, UK, USA, Uzbekistan and Vietnam, suggesting that interest in assessment literacy at the local level for local populations is now widespread and continues to grow. There are undoubtedly other parts of the world where research and scholarship into LAL is well under way but has yet to be published and brought to the attention of the international community; and there may be other regions where LAL is still in the early stages of awareness and development. It might be particularly interesting and instructive, for example, to learn more about LAL experiences in other parts of Asia, in the Middle East and across the many and varied countries of Africa. The opportunities and challenges associated with assessing less commonly taught or spoken languages, with community/heritage languages and with indigenous languages are other areas that it would be good to see represented in future publications. In an increasingly globalised world, other topics that might merit attention include LAL as it relates to other occupational domains (e.g., assessing internationally trained healthcare professionals, or personnel for the call-centre industry) or to sociopolitical contexts (e.g. assessment for citizenship and migration purposes).

Concluding remarks

It is now more than a decade since that early symposium on LAL took place at LTRC 2011 in Ann Arbor. At that time, my discussant remarks highlighted the need for more research to be undertaken and published in the key areas of: LAL construct definition; better communicative engagement with stakeholders to understand LAL needs; and the developmental trajectory of LAL. Although much has been achieved in these areas, there remains much still to be explored and understood. Growing attention on the part of assessment professionals to matters of equity and social justice in their field means that LAL must now embrace a better understanding of how to respond appropriately to test-takers who have special needs (e.g. those with disabilities or in complex circumstances), as well as how to address the negative legacy of

colonialism in assessment practice (e.g. as expressed through so-called ‘native speaker’ linguistic standards), and the privileging of colonial languages (e.g. English and French) to the exclusion of endemic languages in gatekeeping assessments. Furthermore, as society continues to change and as technology continues to advance ever more rapidly, new questions and fresh challenges will emerge with implications for our LAL theory and practice. Among other things, these are likely to include issues around multilingualism and translanguaging within education and society, as well as concerns associated with digital assessment and the role of artificial intelligence.

Albert Weideman’s reflection (Chapter 15) provides a convenient overview of the growth of interest in LAL over the past two decades as he considers how LAL has come to be conceptualised and characterised over that time. Drawing on insights from the philosophy of technology, he evaluates the historical contribution of assessment literacy in terms of a number of gains: these include ethical, economic, juridical, lingua-social, and what Weideman describes as certitudinal gains – associated with the need for appropriate humility on the part of assessment professionals and acknowledgement of our limitations. His framework and analysis will provide readers with some interesting food for thought and reflection. Weideman’s own conclusion is that LAL has brought about a number of advances in the professional understanding of language testing and assessment and has opened up new professional avenues for language test designers and researchers.

In editing these two volumes and bringing them to publication, our hope is they may play a positive role in continuing to advance professional understanding of LAL in our field as well as open up new avenues for LAL-related research and scholarship in the years to come.

References

- Alderson, J C and Hamp-Lyons, L (1996) TOEFL preparation courses: a study of washback, *Language Testing* 13 (3), 280–297.
- Alderson, J C and Wall, D (1993) Does washback exist? *Applied Linguistics* 14, 115–119.
- Alderson, J C, Clapham, C and Wall, D (1995) *Language Test Construction and Evaluation*, Cambridge: Cambridge University Press.
- Bachman, L B (1990) *Fundamental Considerations in Language Testing*, Oxford: Oxford University Press.
- Bachman, L B and Palmer, A S (1996) *Language Testing in Practice*, Oxford: Oxford University Press.
- Bachman, L B and Palmer, A S (2010) *Language Assessment in Practice*, Oxford: Oxford University Press.
- Baker, B (2019) *Multilingual competence, language assessment literacy, and social justice: Tendances dans le domaine d’évaluation de la langue*, plenary talk at the Annual Conference of the Canadian Association of Applied Linguistics [Association canadienne de linguistique appliquée], Vancouver, June 2019.

- Baker, B (2021) *Delineating the construct of LAL: Cultivating professional competency development*, LALSIG Symposium: Language Testing Research Colloquium, International Language Testing Association annual online conference, May 2021.
- Baker, B and Riches, C (2018) The development of EFL examinations in Haiti: Collaboration and language assessment literacy development, *Language Testing* 35 (4), 557–581.
- Beynen, T (2023) *The development of students' assessment literacies as they transition to university: An exploratory case study*, unpublished PhD thesis, Carleton University, Canada.
- Chalhoub-Deville, M and O'Sullivan, B (2020) *Validity: Theoretical Development and Integrated Arguments*, Sheffield: Equinox Publishing.
- Davies, A (1990) *Principles of Language Testing*, Oxford: Basil Blackwell.
- Davies, A (2008) *Assessing Academic English: Testing English Proficiency 1950–1989 – The IELTS Solution*, Studies in Language Testing Volume 23, Cambridge: UCLES/Cambridge University Press.
- Hamp-Lyons, L (1997) Washback, impact and validity: ethical concerns, *Language Testing* 14 (3), 295–303.
- Hawkey, R (2004) *A Modular Approach to Testing English Language Skills: The Development of the Certificates in English*, Studies in Language Testing Volume 16, Cambridge: UCLES/Cambridge University Press.
- Henning, G (1987) *A Guide to Language Testing: Development, Evaluation, Research*, Harlow: Longman ELT.
- Hughes, A (1989) *Testing for Language Teachers*, Cambridge: Cambridge University Press.
- Inbar-Lourie, O (2008) *Language assessment culture*, in Shohamy, E (Ed) *Language testing and assessment Volume 7*, in Hornberger, N (General editor) *Encyclopedia of Language and Education* (Second edition), New York: Springer Science and Business Media, Inc., 285–300.
- Jeong, H (2013) Defining assessment literacy: Is it different for language testers and non-language testers?, *Language Testing* 30 (3), 345–362.
- 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.
- Lado, R (1961) *Language Testing: The Construction and Use of Foreign Language Tests – A Teacher's Handbook*, New York: McGraw-Hill.
- Lave, J and Wenger, E (1991) *Situated Learning: Legitimate Peripheral Participation*, Cambridge: Cambridge University Press.
- Malone, M E (2008) Training in language assessment, in Shohamy, E (Ed) *Language testing and assessment Volume 7*, in Hornberger, N (General editor) *Encyclopedia of Language and Education* (Second edition), New York: Springer Science and Business Media, Inc., 225–240.
- Malone, M E (2013) The essentials of assessment literacy: Contrasts between testers and users, *Language Testing* 30 (3), 329–344.
- 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.
- Rea-Dickins, P (1997) So, why do we need relationships with stakeholders in language testing? A view from the UK, *Language Testing* 14 (3), 304–314.

- Read, J (2000) *Assessing Vocabulary*, Cambridge: Cambridge University Press.
- Saville, N (2000) Investigating the impact of international language examinations, *Research Notes* 2, 4–7.
- 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.
- Schön, D (1983) *The Reflective Practitioner – How Professionals Think In Action*, London: Temple Smith.
- Stiggins, R J (1991) Assessment literacy, *Phi Delta Kappan* 72 (7), 534–539.
- Stiggins, R J (1995) Assessment literacy for the 21st century, *Phi Delta Kappan* 77 (3), 238–245.
- Stiggins, R J (2014) Improve assessment literacy outside of school too, *Phi Delta Kappa International* 96 (2), 67–72.
- Stoyonoff, S and Chapelle, C A (2005) *ESOL Tests and Testing: A Resource for Teachers and Program Administrators*, Virginia: TESOL Publications.
- Taylor, L (2000) Stakeholders in language testing, *Research Notes* 2, 2–4.
- Taylor, L (2009) Developing assessment literacy, *Annual Review of Applied Linguistics* 29 (1), 21–36.
- Taylor, L (2013) Communicating the theory, practice and principles of language testing to stakeholders: Some reflections, *Language Testing* 30, 403–412.
- Taylor, L (2023) Reframing the discourse and rhetoric of language testing and assessment for the public square, *Language Testing* 40 (1), 47–53.
- Taylor, L and Green, A (2020) The role of academic institutions in language testing research and consultancy, in Taylor, L and Saville, N (Eds) *Lessons and Legacy: A Tribute to Professor Cyril J Weir (1950–2018)*, Studies in Language Testing Volume 50, Cambridge: UCLES/Cambridge University Press, 175–206.
- Taylor, L and Harding, L (2020) *A testing time for testing: Assessment literacy as a force for social good in the time of coronavirus*, available online: acss.org.uk/a-testing-time-for-testing-assessment-literacy-as-a-force-for-social-good-in-the-time-of-coronavirus/
- Tsagari, D (2020) *Language Assessment Literacy: From Theory to Practice*, Newcastle: Cambridge Scholars Publishing.
- Weigle, S C (2002) *Assessing Writing*, Cambridge: Cambridge University Press.
- Weir, C J (1990) *Communicative Language Testing*, Hoboken: Prentice Hall.
- Weir, C J (1993) *Understanding and Developing Language Tests*, Hoboken: Prentice Hall.
- Weir, C J, Vidaković, I and Galaczi, E D (2013) *Measured Constructs: A History of Cambridge English Language Examinations 1913–2012*, Studies in Language Testing Volume 37, Cambridge: UCLES/ Cambridge University Press.
- Wenger, E (1998) *Communities of Practice: Learning, Meaning and Identity*, Cambridge: Cambridge University Press.

Section 1

Empirical Research Studies

2

Developing learning-focused feedback practices: An evaluation of English for Academic Purposes teacher feedback literacy at a Sino-foreign university

Joseph Arthur Davies

Duke Kunshan University, People's Republic of China

This chapter explores the feedback literacy of Higher Education (HE) English for Academic Purposes (EAP) teachers based at a Sino-foreign joint venture university. Specifically, the chapter aims to:

- Critically review the suitability of current teacher feedback literacy research for those teaching Confucian Heritage Culture learners within Sino-foreign HE contexts
- Analyse a small corpus of interviews with Sino-foreign HE EAP practitioners about their feedback beliefs, design, and practices
- Map the results to a newly developed Sino-foreign teacher feedback literacy framework
- Discuss the implications of the results for HE EAP teachers wishing to develop their own teacher feedback literacy

Introduction

It is widely accepted that the provision of high-quality feedback can have one of the biggest influences on student achievement (Hattie and Timperley 2007) within Higher Education (HE). Carless and Boud (2018:1,315) define feedback as ‘a process through which learners make sense of information from various sources and use it to enhance their work or learning strategies’. Such a definition views feedback as more than just information transmitted from the teacher. Instead, it highlights students’ active role in the feedback process through sense-making. Despite the obvious benefits of quality feedback for student learning and development, the actual impact feedback has upon

students is highly variable. For example, Carless, Salter, Yang and Lam (2011) found that feedback on end-of-course summative assignments had little impact on students' future development, and Robinson, Pope and Holyoak (2013) showed that students can lack the skills required to interpret feedback comments. Moreover, feedback is a commonly cited factor that university students report being unhappy with according to student course evaluation data (Carless and Boud 2018:1,315), which is mostly quantitative and gives little insight into the source of such dissatisfaction. In addition, teachers often report frustration with the time it takes to produce feedback and the limited obvious impact it has on students (Price, Handley and Millar 2011). Therefore, although much feedback research has been conducted in HE contexts (Tight 2021), improved understanding of effective feedback is needed.

Of particular interest to current feedback researchers is the concept of feedback literacy and how this literacy requires learner training and development in order to maximise feedback gains for learners. As assessment and feedback are often inextricably linked through both policy and practice, feedback literacy can be considered an important strand of the overarching assessment literacy domain. Therefore, it is appropriate to bring research and discussion pertaining to feedback literacy to the attention of assessment literacy scholars and practitioners so that mutual insights can be shared. However, despite much recent research on student feedback literacy (Carless and Boud 2018, Molloy, Boud and Henderson 2020), Carless and Winstone (2023) point out that research on teacher feedback literacy is lacking. Although research on student feedback literacy can help practitioners understand students' roles in the feedback process, it is teachers themselves who must set up the optimum feedback conditions for success. This is particularly important for those teaching within Sino-foreign HE contexts that involve collaborations between a Chinese and foreign university at programme, college or university level. This is because Confucian Heritage Culture learners¹ have to adapt to completely new, and often contradictory, social and educational systems while studying in English as a Medium of Instruction (EMI) in a foreign language (Ou and Gu 2021), often for the first time.

EMI is increasingly commonplace in East and South East Asia as a result of the 2012 Asia-Pacific Economic Cooperation (APEC) agreement, which sought to encourage movement within HE in the region (Kirkpatrick 2017). It is also a common feature of Transnational HE (TNHE) programmes

¹ Learners originating from a culture that is underpinned by the teachings of Confucius (e.g. P.R. China) that promote benevolence, doing what is right, and loyalty to one's superiors. One respondent summarised this as 'the teacher's the only one that has the knowledge ... what the teacher says matters and what classmates say doesn't really matter as much because they don't know'.

that enroll students located in a country different from where the awarding institution is based. However, students in Asia (Kirkpatrick 2017:32) and China (Hu and Lei 2014) have been shown to lack the linguistic ability to successfully study in English. Despite this, there are 2,539 TNHE programmes in China (Ou and Gu 2021), demonstrating huge demand for international education in English. Consequently, if feedback can have one of the largest impacts upon HE learners (Hattie and Timperley 2007), yet students cannot fully understand feedback in English, perhaps a reconsideration of teacher feedback literacy is required for Sino-foreign HE contexts.

The interdisciplinary field of English for Academic Purposes (EAP) provides the ideal site to examine feedback as its practitioners help students manage the transition to university through developing academic writing (and other academic literacies) and by providing individualised attention and feedback (Hyland 2019). Specifically, the research reported in this chapter aims to evaluate the feedback beliefs, designs, and practices of EAP teachers based at a Sino-foreign joint venture university by addressing the following research questions (RQs):

RQ1: To what extent do EAP teachers' beliefs about feedback demonstrate teacher feedback literacy?

RQ2: To what extent do EAP teachers' reported feedback designs and practices demonstrate teacher feedback literacy?

Literature review

Carless and Boud (2018) take a social constructivist view of feedback being generated through dialogue, sense-making and the co-construction of knowledge between teachers and students. However, as teachers and students often have differing perceptions of what feedback actually is (Carless 2006, Dawson et al 2019), co-constructing such knowledge may prove challenging. For example, Adcroft (2011) found that teachers and students held different views of feedback leading to dissonance. This is supported by Hyland (2019) who interviewed university students from a range of disciplines in Hong Kong about the kinds of messages they took from feedback. Hyland (2019) concludes that greater alignment is required between teacher and student expectations of feedback, which should be explicitly clarified at the start of a course. Therefore, developing teacher feedback literacy so that feedback practices actually complement the needs of learners is important.

Carless and Winstone (2023:153) define teacher feedback literacy as 'the knowledge, expertise and dispositions to design feedback processes in ways which enable student uptake of feedback and seed the development of student feedback literacy'. This definition emphasises how teachers should facilitate

learners' use of feedback, which ultimately requires a shared understanding of what feedback actually is. Previous research has shown that, traditionally, teachers perceive feedback as information transmission, with minimal impact on students' actual uptake (Jiang, Yu and Zhao 2019). In contrast, more recently there has been a scholarly trend to view feedback as dialogic communication between teachers and students (Ajjawi and Boud 2018, Carless 2020, Carless and Boud 2018, Winstone and Carless 2020) with the aim of supporting students' use of feedback to improve academic performance. Yet, the extent to which such current thinking has been applied by teachers to their feedback practice remains unclear. For example, Jiang and Yu (2021) recently found that Chinese EFL teachers had reverted back to information transmission modes of feedback during the COVID-19 pandemic.

Based on a review of the literature, Carless and Winstone (2023) create a tripartite framework of teacher feedback literacy that aims to help teachers improve their feedback practice comprising design, relational and pragmatic dimensions. Importantly, all three dimensions can be enhanced with technology as shown by Nicol, Thomson and Breslin (2014), who discuss the benefits and affordances of using *Turnitin Feedback Studio*² for peer feedback. More recently, Boud and Dawson (2021) propose an empirical teacher feedback literacy framework consisting of macro, meso and micro levels. However, in a separate publication, Winstone and Carless (2020:10) acknowledge that 'the design of feedback processes takes place within a complex interaction of intrapersonal, interpersonal and contextual influences'.

Winstone and Carless' (2020:10) comment highlights the fact that teacher feedback can be inextricably linked to student feedback engagement. Jönsson's (2013) meta-analysis identified five key challenges preventing student feedback engagement: usefulness, detail and specificity, emotion, lack of feedback strategies, and linguistic inability to understand feedback. To support this, Winstone, Nash, Rowntree and Parker's (2017) empirical research identified four additional challenges: awareness, cognisance, agency, and volition. Therefore, feedback literate teachers will need to adopt strategies to overcome such challenges. For instance, linking to the above-mentioned complex interactions involved in feedback design (Winstone and Carless 2020:10), teachers could embed elements of Winstone, Nash, Parker and Rowntree's (2017) framework and taxonomy within their feedback design. These elements include self-appraisal, assessment literacy, goal-setting and self-regulation, and engagement and motivation (SAGE). However, they admit such skills are complex and students require regular opportunities to practise and hone their ability to successfully engage with feedback.

2 An internet-based software program that checks students' academic submissions for text similarity and facilitates teachers and peers to provide feedback.

Crucial to second/foreign language HE teaching contexts is Jönsson's (2013) finding that students fail to engage with academic feedback language due to a linguistic inability to understand. This is supported by Higgins (2000:1) who argues 'many students are simply unable to understand feedback comments and interpret them correctly'. Carless (2006) goes further to explain that feedback is provided within academic discourse to which many students, especially those adapting linguistically, academically and socioculturally to new Sino-foreign HE contexts, simply do not have access. Therefore, this linguistic inhibitor is exacerbated within Sino-foreign HE teaching contexts for learners who are not only adapting from one socioeducational background to another, but are also studying in a foreign language, often for the first time (Ou and Gu 2021). Despite attempts within the research literature to improve understandings of teacher feedback literacy, the extent to which such findings apply to Sino-foreign HE contexts is unclear.

Methodology

Research design

This study aims to build upon the current theoretical knowledge about teacher feedback literacy by exploring the actual feedback beliefs and reported designs and practices of EAP teachers at a Sino-US joint venture university. It is hoped that this research can lead to the creation of a new understanding of teacher feedback literacy that is more appropriate for HE contexts where Confucian Heritage Culture learners, and other second language learners in general, are engaged with English curriculum, pedagogy and assessment.

As the university forming the focus of this research is a relatively new Sino-US joint venture university, the interdisciplinary liberal arts undergraduate curriculum is still under development and so requires ongoing evaluation and refinement. Therefore, this research aims to contribute to the development of first year undergraduate EAP curriculum and pedagogy by performing a formative evaluation of ongoing EAP teacher feedback practices to enable improved programme performance. In order to operationalise teacher feedback literacy within this context, after taking inspiration from Carless and Boud (2018) and Carless and Winstone (2023), an eight-point framework of the following competences was developed for comprehensibility and practicality:

1. Create the ideal feedback environment
2. Engage in feedback dialogue
3. Encourage students to make judgements about academic work

4. Encourage students to take action on feedback information
5. Share feedback responsibilities
6. Help learners understand feedback
7. Be pragmatic and efficient
8. Engage with feedback literature

In order to facilitate the implementation of this formative evaluation in terms of the relationships between planned resources and activities and intended short, medium and long-term results (Hayes, Parchman and Howard 2011), a logic model based on Wyatt Knowlton and Phillips (2013:10) was devised to guide the process. Additionally, as recommended by Saunders (2000:9), the RUFDATA model was used to facilitate the evaluation design as it ‘involves a process of reflexive questioning during which key procedural dimensions of an evaluation are addressed, leading to an accelerated induction to key aspects of evaluation design’. Using such models allowed outcomes, impacts, and outputs to be carefully planned before working backwards to identify the required activities and resources needed to achieve them. This also helped plan appropriate evaluation outputs, which directly impact usability of evaluation results, and facilitated consideration of the most appropriate vehicle of communication to inform programme users (i.e. EAP teachers) of these results. For instance, the evaluation outputs for this research included the design of internal faculty feedback workshops and feedback practice guidelines intended to act as bridging tools (Saunders 2012:429) to improve teachers’ feedback literacy.

Interviews

As the goal was to evaluate not only the EAP faculty’s reported feedback practices, but also their rationale for such practices, this research adopted a qualitative approach to data collection and analysis using interviews with current EAP faculty members. Specifically, 45-minute semi-structured interviews were conducted with seven conveniently sampled EAP teaching faculty members at the Sino-foreign university in question. Seven interviews were deemed sufficient as this accounted for over 50% of the EAP teaching faculty at the time. Interviews were conducted online in English, facilitated by the author and recorded for accuracy and later transcription. Questions were focused on three main topics – EAP teachers’ feedback beliefs, design, and practice – and interviews were kept intentionally dialogic and flexible.

To supplement self-reports pertaining to teachers’ feedback practices and to promote recall, participants provided example feedback documentation representing their EAP feedback practices in advance of the interview. For example, respondents provided marked-up student essays, completed feedback forms, and peer feedback classroom task instructions. This was

also shared during the interview to encourage participants to explain their rationale for such practice. Such a research design was successfully used by Jiang and Yu (2021) who asked interviewees to provide snapshots of their digital footprint when investigating EFL teachers' feedback practices during the COVID-19 pandemic. They argue that such snapshots help to situate teachers' self-reported data into more concrete recollections. Prior to data analysis, in an attempt to increase validity, interview transcripts were returned to participants to check accuracy, clarify any points, add further detail, and confirm identifiable information had been suitably redacted via member checks, where participants are invited to read and validate interview transcripts or results, as recommended by Maxwell (2004).

Interview analysis

Following Braun and Clarke's (2006) model for conducting thematic analysis, interview data were reviewed to note salient themes. Once member-checked, the data were deductively coded thematically using NVivo 12 software. Deductive analysis was conducted as prior research on teacher feedback literacy was consulted in an attempt to develop a framework to which the results could be aligned. This approach also increased sensitivity to subtle and relevant features within the data (Tuckett 2005). Although qualitative researchers generally agree that there is no specific set number of interviews that can be assumed to achieve saturation (Morse 1995), by the seventh interview no new themes were extracted from the data. Finally, it is important to note the dual positionality of the researcher and analyst as both a cultural member and cultural commentator (Braun and Clarke 2006:94), and how the researcher brings their own knowledge, assumptions and expectations that actively contribute towards qualitative data analysis. This can be helpful when the researcher has intimate knowledge of the context that can be used to better understand and interpret respondents' comments.

Results

As the interviews were designed around three major topics (feedback beliefs, design, and practices), these topics will be used to broadly organise the results before going into more detail of the salient themes within each topic. Pseudonyms have been used throughout to refer to the EAP teachers.

EAP teachers' feedback beliefs

The sampled EAP teachers believe the purposes of feedback relate to supporting student improvement, providing a meaningful response, sharing relevant information, and facilitating student learning, as illustrated by

the following comments. For example, Charlotte described the purpose as: ‘providing information on work that was done by students in order to ... “improve” ... improve whatever it is that they’re working on.’

Furthermore, Sarah highlighted the challenge she faces when balancing between grade justification and providing formative feedback for improvement: ‘... having to give it a grade forces me to have to sometimes justify things that are sometimes just a feeling or an intuition ... trying to explain to a student, why they got a B+ and another student got an A- ... and really what the student should be focusing on is just my feedback as a reader and trying to make the paper better.’

The teachers also believe that conversations, reflection, drafting and peer feedback are effective methods of promoting feedback. Beliefs about conversational modes of feedback were well summarised by Andrew who stated: ‘... feedback is really just an ongoing conversation ... you give them feedback on a paper, they ask you questions. You be more specific, they offer something else then make some changes.’

Mary emphasised that feedback conversations can take place with others beyond the teacher: ‘... the goal is for it to really be kind of a conversation, but for the students to be able to direct that conversation between themselves and their peers and their teachers and other people that they’re wanting to get feedback from.’

Moreover, George highlighted the time-saving benefits of peer feedback as well as the importance of the drafting process and how peer feedback can support this: ‘... it promotes the writing process ... this isn’t a test, where you have 45 minutes ... You have the time and space ... to really go over things ... peer feedback promotes the draft system.’

Participants also believe that responsibility for the feedback process should be shared between teachers and students. This was highlighted by Simon who reported: ‘I think everybody’s responsible for it ... for feedback to work, both sides need to be interested in the feedback and sort of engaged in the feedback.’

Teachers reported their belief in taking personal responsibility for scaffolding, counselling, creating a positive environment, designing feedback processes, providing information on academic work, and evolving their own feedback beliefs. Mary outlined teacher feedback responsibilities as follows: ‘By the way we give feedback and then talking with students about feedback, and having them practice asking for feedback, having them practice giving each other feedback, having them reflect on how they’re using the feedback as they’re revising or in their learning.’

Teachers also noted the influence of different teaching contexts and student demographics upon feedback practices. For example, Theresa reflected on how Confucian Heritage Culture students’ past educational experiences can negatively impact peer feedback as they focus only on what the teacher says and not on their peers’ comments.

EAP teachers' feedback design

The results show that teachers are able to demonstrate awareness of how EAP course design can support their feedback beliefs. For example, Andrew explains how he designs his course around a drafting process to facilitate feedback conversations and how oral presentations are designed to encourage oral feedback and reflection: '... there's two major veins of feedback and one is feedback on papers, ... we have three drafts of every paper ... for the second paper that we do, that feedback process is not just written ... they give presentations at the end of the semester before their final papers are due, and I really encourage folks to consider those presentations to be part of this feedback process.'

Furthermore, Mary designs her EAP course around a portfolio-based drafting process that includes peer feedback and writer autonomy as students select which drafts to improve and submit. She also encourages student reflection through interactive cover sheets: '... they could choose two of those short writing assignments to revise and submit ... with a cover sheet for each of the revisions talking about what the assignment purpose was, what they're trying to accomplish in the piece of writing, how they revised it, why they chose it ...'

As well as encouraging students to reflect on feedback in the design process, it was evident that participants themselves reflected upon their own practice. For instance, Theresa reported: 'You know just once I think I've figured it out, I have a different batch of students ... I'm going to adjust it a little bit ... So I'm always playing with it and adjusting.'

The sampled EAP teachers also suggested improvements to current feedback practices, including increasing class time, raising students' awareness of and confidence with feedback, developing a feedback orientation, developing a student feedback community, and developing a student feedback toolkit of resources.

EAP teachers' reported feedback practice

Conversations

Linking very closely to the previous feedback beliefs, the most commonly reported feedback practice related to feedback conversations and encouraging dialogue throughout the feedback process. For instance, Charlotte mentioned benefits to feedback conversations that extend beyond the provision of information to support teachers' understanding of students' academic writing intentions, implying a clarification function: '... it's a benefit to the instructor to understand what the student is writing ... maybe you didn't understand what they were saying. So get the conversation going with them ... nine times out of ten, students can explain it well.'

Peer feedback

All EAP teachers reported using peer feedback, supporting the previously identified beliefs about its benefits. For example, Mary stated that she was a firm proponent of peer feedback and demonstrated reflexive practice when she realised her students may lack the linguistic ability to provide appropriate non-offensive peer feedback. In response, she discussed how she developed detailed feedback guides with sample language, explaining the benefits as ‘relieving face-threatening pressure’ as students use the teacher’s model language to critique peers.

Mary also takes inspiration from the feedback literature to facilitate peer feedback circles in her classroom, where pairs peer review each other’s work and then join another pair to orally summarise their review. Mary justifies this approach as being ‘really valuable for the writer to hear, and a good way to focus what that reviewer was going to say’. Theresa echoed this sentiment in response to students not valuing peer feedback: ‘I make it clear to them ... research shows that the one that’s giving the feedback is actually often benefiting more than the one that’s receiving the feedback ... Having that whole speech with them really gets buy in with my students.’

Theresa also mentioned the challenges of adopting peer feedback with Confucian Heritage Culture learners, explaining how they lack expertise in evaluating others’ work and so require it to be modelled: ‘... they don’t really have guidance when it comes to peer feedback, so they don’t know what they’re doing ... I have to kind of model feedback for them before I get them into the peer review.’

Emotional awareness

Participants were also aware of the emotional impact feedback can have upon their students. Emotional awareness was categorised as personalising feedback, building relationships and providing positive comments. For example, Theresa pays close attention to her students’ needs: ‘I try to the best of my ability to adjust my feedback, based on what is most useful for different students.’

Charlotte also personalises her feedback but points out how small class sizes afford her this luxury: ‘It’s about having as the instructor some sensitivity to individual differences and needs in order to help the student. But I can do that at our university because I have 12 students.’

Interestingly, the EAP teachers held conflicting views about how critical feedback might impact students emotionally, with some teachers being more direct than others. For instance, Simon explained that he does not need to take students’ emotions into consideration when providing feedback as he feels as long as a positive relationship is built, his Confucian Heritage Culture students are able to accept direct criticism as they are used to it from high school: ‘... in our context here in China, I think we’re in a good situation

where we don't have to worry about [students' emotions] too much. But as much as I try to be an approachable and friendly teacher and I want to have a good relationship with them, I do think feedback needs to be tough sometimes.'

In contrast, Theresa feels that her Chinese students find it hard to deal with direct criticism of their academic work: '... they get a bunch of feedback on their writing, which suggests that they're not as great as they think they are. And it's harsh. And they remember it ... But I really try and consider their feelings, and mostly I do that by trying to develop a positive relationship with the student. So they see that my goal is to help them be better, it's not to point out that they're wrong.'

Learner agency

Additionally, the EAP teachers made direct connections between their feedback practices and promoting learner agency. Key themes included student reflection, student motivation, writer autonomy and preparation beyond EAP courses. For example, Mary explained how she feels responsible for preparing students for the following academic year when they will no longer benefit from EAP courses and the extensive feedback provided: 'I feel like I need to get them ready for next year, when they're not going to be in an EAP course and just getting loads of feedback.'

Furthermore, George utilises feedback conversations as an opportunity to promote writer autonomy by reinforcing student agency in the decision-making process: 'I want it to be a dialogue. Because the one thing I do firmly believe in ... you're the writer ... this is your piece of writing, you own it, I can give you suggestions ... you make the decision on this, and that's where the dialogue does come in.'

Uptake

All EAP teachers mentioned positive instances of student uptake. For instance, Charlotte reported being 'really impressed in general at how well students pick up feedback' and George reported: 'The majority of my students do use my feedback. Because I see that the mistakes they made in their first or second paper and then in the third and fourth papers they're not making those mistakes, they're taking my advice, they're really making changes.'

Mary specified how 'stronger' students engage more with the feedback process: 'They seem to be putting more time into their drafting, they seem to be formulating better questions and requests for feedback, and they seem to be using the feedback more.'

However, despite promoting writer autonomy in selecting which feedback to act upon, she felt her students were not making the best choices: '... some of them are picking and choosing the feedback that they use. But of course

that's what I want them to do, right? They own the writing. But they're not always making what I would consider to be the best decisions in terms of the improvements of their writing.'

All EAP teachers reported a lack of student uptake, because as Andrew commented, 'of course there'll be some students that don't listen'. He continued to provide a specific example: '... this one student did this presentation today, and it was obviously just based on a first draft, which I gave comprehensive feedback on. Obviously this student did not take into account anything that I had written.'

When teachers were asked to explain the lack of student uptake, responses centered on students' time pressure, heavy assessment load, and low motivation for mandatory EAP courses.

Technology

Respondents reported different levels of technology adoption within their feedback practices, ranging from simple written comments on paper scripts and face-to-face meetings to screencasted audio-video feedback, a digital recording of a computer screen output, and integrated *Turnitin* *Grademark* feedback. One teacher who embraced technology was Sarah who regularly uses video screencasting to provide individualised voiceovers of her feedback process in an interactive manner that she believes can reduce misunderstandings: '... reading the paper and recording myself talking through it, because I think some of the issues with feedback I've found ... are our students understanding what the feedback is ... where I'm like "I wrote on your paper such and such" and they're like "Oh, I thought this meant ...".'

Time and space

All the sampled EAP teachers reported challenges associated with the time it takes to design and provide feedback in their EAP courses. For instance, Charlotte admitted: 'I think I give too much feedback and I take too much time.' Sarah supported this sentiment as she reported: '... the major challenge is time. Having enough time to give the kind of feedback I would want to be able to give.'

Teachers did, however, report adopting different strategies to improve efficiency. For example, Sarah uses the previously mentioned technology to increase efficiency: 'Something I could say, and it takes me three seconds, but if I had to type it out, it would just be so much more time ... since I'm screen sharing I could actually just type something into Google and show the students the thing.'

Discussion

In general, the reported EAP teacher feedback beliefs and reported practices do demonstrate teacher feedback literacy as they align well with what Winstone and Carless (2020) term new paradigm learning-focused feedback practices. For example, they follow dialogical characteristics as reported by Nicol (2010:505), and involve ‘a dialogical and contingent two-way process that involves coordinated teacher-student and peer-to-peer interaction as well as active learner engagement’. The reported feedback practices promote teacher-facilitated, peer, and inner dialogue, for example, via interactive cover sheets as successfully demonstrated by Bloxham and Campbell (2010). Overall, the reported feedback beliefs and practices align well with current social constructivist perspectives on feedback (Carless and Boud 2018), something that is well demonstrated by George’s use of dialogue in the feedback process to offer students suggestions but ultimately remind them of their own agency to make final decisions on their work.

Furthermore, the sampled EAP teachers believe that student and teacher responsibilities within the feedback process should be shared. This also corresponds with a social constructivist approach to feedback which holds the view that individual and shared understandings are co-constructed via dialogue and sense-making (O’Donovan, Rust and Price 2016). However, few reported that teachers should be responsible for providing students with strategies to take positive action on feedback despite Nash and Winstone (2017) arguing that this is a key responsibility. This may be because the EAP teachers focused more specifically on students’ academic writing and language development rather than improving their feedback literacy. Consequently, the sampled EAP teachers may need to consider further how their feedback design and practice can better support students to take positive action on feedback information.

EAP teachers also explicitly discussed the impact Confucian Heritage Culture students’ previous educational experiences may have upon successful feedback practices. This is important within Sino-foreign HE contexts, where such learners must quickly adapt to completely different social, educational and linguistic norms (Ou and Gu 2021). As previous education has been shown to impact feedback expectations (O’Donovan 2017), students’ transition to Sino-foreign HE contexts could inhibit the successful implementation of social constructivist feedback practices. For instance, EAP teachers reported students lacking the linguistic ability to both understand feedback and provide appropriate peer feedback without offending classmates. This supports previous research showing how students’ linguistic inabilities can inhibit both the feedback process (Higgins 2000, Jönsson 2013) and students’ ability to successfully study EMI courses (Hu and Lei 2014). Such observations are also supported by Zhang and Head

(2010:3) who argue the concept of *face*, a complex sociocultural phenomenon linking to ideas of honor, dignity and self-worth, can discourage Chinese students within international or TNHE contexts from disagreeing with others. Reported successful strategies to overcome such challenges included scaffolding students through modelling, providing students with functional example peer feedback language, and emphasising the benefits of peer review for the reviewer as opposed to the receiver as evidenced by Lundstrom and Baker (2009).

Despite teachers reporting instances of limited student uptake, this was not necessarily attributed to students misunderstanding feedback or to the key challenges outlined by Jönsson (2013) or Winstone, Nash, Rowntree and Parker (2017). Instead, teachers hypothesised that students' academic pressure, heavy workloads, poor time management and previous educational experiences were to blame. However, perhaps an opportunity has been missed, during the widely reported feedback conversations, to obtain information directly from students about why they failed to engage with feedback. For example, time during scheduled feedback tutorials could easily be used to explore why students did not use feedback during the writing process. Further exploration of such reasons could provide rich insight into student uptake within Sino-foreign contexts and beyond.

Another point raised by participants was the importance of emotional awareness within the feedback process. This is supported by Carless and Winstone's (2023) relational dimension of their teacher feedback literacy framework and has been found to be a key barrier to student engagement with feedback (Jönsson 2013). It was interesting that some EAP teachers held opposing views about the emotional impact their feedback can have upon learners, with some hedging criticism while others were much more direct. This issue is discussed by Hyland and Hyland (2001) who suggest that when teachers hedge their criticisms and suggestions, albeit to protect their students, feedback may be misinterpreted by learners. One point of agreement was the need to foster a positive relationship with learners in order for feedback to be successful. Ultimately, more follow-up research is needed to ascertain from students themselves how the emotional impact of feedback can affect their engagement with it and their subsequent academic development.

As noted by Carless and Winstone (2023), technology-enabled feedback is a key component of contemporary feedback practices, and feedback literate teachers should be able to use technology to improve feedback (Boud and Dawson 2021). Respondents reported different levels of sophistication in terms of their technology-enhanced feedback practices. For example, one respondent adopts a similar approach to that reported by Mahoney, Macfarlane and Ajjawi (2019) by providing talking head and screencast combination video feedback. Other respondents view *Turnitin Grademark* as

a useful feedback tool and interface as opposed to simply a punitive deterrent against plagiarism, supporting Kostka and Maliborska's (2016) research. However, in general, the results show that EAP teachers could further utilise technology to improve their efficiency. As all respondents reported challenges associated with the time it takes to provide high-quality feedback, echoing much of the feedback literature (Price et al 2011), perhaps technology could be better used to overcome such pragmatic challenges, as recommended by Carless and Winstone (2023).

Finally, despite Hyland's (2019) assertion that greater alignment is required between teacher and student expectations of feedback, and that this should be explicitly set out at the start of a course, this was not reported by the sampled EAP teachers. Perhaps more time needs to be allocated at the start of courses to set out feedback purposes, expectations and roles, much like what is routinely done when introducing course syllabuses and assessment strategies. Despite this, the results do seem to align with what Nicol (2010) describes as a tutorial feedback system, often adopted by a number of select universities. However, he stresses that due to general trends of increased HE student recruitment and the pressure this puts on teachers, such feedback practices are rarely sustainable. The wider application of the feedback practices reported will depend on readers' specific educational contexts and in particular their class sizes and teaching loads.

Sino-foreign teacher feedback literacy framework and key recommendations

In order to form a comparative model to which the results from this particular context could be aligned, a new teacher feedback literacy framework was devised that is relevant for Sino-foreign HE teaching contexts (See Table 1). The framework, based upon both the feedback literature and results from this study, also provides key recommendations for language teachers wishing to enhance their own teacher feedback literacy within contexts beyond Sino-foreign HE settings. Inspiration was taken from Carless and Boud's (2018) and Carless and Winstone's (2023) research to acknowledge the interplay between student and teacher feedback literacy. However, whereas Carless and Winstone (2023) compile a general tripartite framework comprising design, relational and pragmatic dimensions, the eight-point framework presented here is intended to be brief and actionable by summarising salient features that language teachers within Sino-foreign HE contexts and beyond should pay attention to and take account of. The eight teacher feedback literacy features have each been broken down into several descriptive criteria to present a list of specific actions. A tick (✓) means that an action was reported by all or most respondents in this study, a dash (-) means it was reported by some respondents, and a cross (X) means it was reported by none or very few respondents.

Table 1 Sino-foreign teacher feedback literacy framework

Teacher feedback literacy feature	Descriptive criteria Key: (✓) Fully represented by the results of the study (-) Somewhat represented by the results of the study (X) Not represented by the results of the study
1. Create the ideal feedback environment	<ul style="list-style-type: none"> • Build positive trusting relationships between teacher, students and classmates. (✓) • Personalise feedback comments for individual student needs. (✓) • Encourage students to request feedback through various channels. (-) • Design tasks that encourage students to discuss feedback and their reaction to it together. (X)
2. Engage in feedback dialogue	<ul style="list-style-type: none"> • Maximise opportunities for face-to-face tutorials to discuss and clarify feedback with students in person. (✓) • Engage in dialogue with peers and colleagues to share feedback tasks, examples, and ideas. (✓) • Engage in teacher–student meta-dialogue about feedback and assessment in general. (X)
3. Encourage students to make judgements about academic work	<ul style="list-style-type: none"> • Design tasks that increase opportunities for peer feedback and evaluation. (✓) • Design tasks that increase opportunities for self-evaluation and reflection. (✓) • Facilitate collaborative student evaluations of academic work. (X) • Reduce students' reliance on information transmission from the teacher. (X)
4. Encourage students to take action on feedback information	<ul style="list-style-type: none"> • Design curricula and linked assessments specifically to show development over time. (✓) • Require students to show and/or explain how they have positively used feedback from previous drafts/assignments. (-) • Emphasise the process and use of feedback, not only the final product, and consider this as an explicit criterion within assessment rubrics. (X)
5. Share feedback responsibilities	<ul style="list-style-type: none"> • Adopt a process-oriented social constructivist view of feedback. (✓) • Promote learner agency and ownership over the feedback process. (✓) • Clarify teacher and student feedback roles, responsibilities, and expectations at the start of courses. (X)
6. Help learners understand feedback	<ul style="list-style-type: none"> • Provide error correction codes with clear examples of correct and incorrect sentences in context. (✓) • Provide model feedback and peer feedback examples and discuss these with students in class. (-) • Provide model functional language that students can use when completing peer feedback tasks. (-)

Table 1 (continued)

Teacher feedback literacy feature	Descriptive criteria Key: (✓) Fully represented by the results of the study (-) Somewhat represented by the results of the study (X) Not represented by the results of the study
7. Be pragmatic and efficient	<ul style="list-style-type: none"> • Utilise technology to improve timeliness, efficiency and portability. (-) • Balance time spent on feedback input with what is actually useful for students. (X) • Make the feedback process satisfying for teachers. (X)
8. Engage with feedback literature	<ul style="list-style-type: none"> • Regularly read academic research related to feedback to keep abreast with current trends. (✓) • Share feedback research with students to facilitate better understanding of its importance to learning e.g. peer feedback and self-assessment benefits. (-) • Engage in action research to share (internally and externally) results (successes and failures) of teachers' own feedback practices. (X)

The intention is that this framework, combined with the results, can be used as a forward-looking tool to indicate areas of teacher feedback literacy that EAP teachers may wish to develop further both within and beyond Sino-foreign HE settings (albeit based on this specific sample and context).

Finally, for readers planning to evaluate feedback practices within their own teaching contexts, the benefits of first designing logic (Hayes et al 2011) and RUFDATA (Saunders 2000) models should not be underestimated. For example, establishing the intended short, mid and long-term outcomes and impacts and associated outputs before working backwards to identify the required activities and resources was extremely helpful. Using these models also helped maintain the evaluation focus and establish appropriate vehicles of communication for dissemination (Saunders 2000, 2012). For example, using the above-mentioned Sino-foreign teacher feedback literacy framework, planned outputs acting as bridging tools included feedback guidelines, faculty feedback workshops and a new feedback policy. Ideally, such outputs will lead to short-term outcomes including faculty reflections on their feedback practice, and longer-term outcomes including an overall change in organisational feedback culture and increase in student and teacher feedback literacy.

Conclusion

Sustained dissatisfaction from both teachers (Price et al 2011) and students (Winstone and Carless 2020:5) of current HE feedback practices calls for

a reconsideration and adoption of more social constructivist learning-focused practices. This is important for EMI Sino-foreign HE contexts, where learners are transitioning to a new educational and linguistic system (Ou and Gu 2021). The research reported in this chapter has attempted to extend current knowledge of teacher feedback literacy by evaluating the extent to which EAP teachers at a Sino-foreign university demonstrated feedback literacy within their reported feedback practices. Through semi-structured interviews exploring their feedback beliefs and practices, results indicated that participants showed an awareness of how to operationalise their feedback beliefs into practice and that, in general, the reported beliefs and practices did demonstrate teacher feedback literacy.

Key themes relating to participants' feedback practices were: conversations, peer feedback, emotional awareness, learner agency, level of uptake, role of technology, and challenges of time and space. The results were also aligned to a newly developed Sino-foreign teacher feedback literacy framework. This eight-point framework was used to identify areas of good practice and areas requiring further development based on the results. Furthermore, the framework provides key recommendations for readers within various HE contexts to enhance their own teacher feedback literacy. However, in order to create a more complete picture, analysis of more cases in other contexts would be beneficial. Additionally, this research only analysed teacher perspectives and reported feedback practices, of which at times assumptions were made of students. Therefore, follow-up research investigating Sino-foreign HE students' feedback perspectives and practices would also be helpful.

To conclude, acknowledging that, rightly or wrongly, assessment and feedback are very often linked through policy and practice, it is worth bringing the feedback literacy discussion to the language assessment literacy table more often. As this research has shown, language, culture, and context are powerful influencers of teacher feedback literacy within foreign language teaching contexts. Therefore, feedback literacy should be viewed as an important strand of the overarching language assessment literacy domain. Finally, it is hoped that raising awareness of the explicit competences required to develop teacher feedback literacy for language teachers will contribute towards a better understanding of language assessment literacy.

Acknowledgements

This research was undertaken as part of the PhD in Educational Research – Higher Education in the Department of Educational Research at Lancaster University. I am pleased to acknowledge the contribution of Dr Gemma Derrick and my peers in supporting the development of this research.

References

- Adcroft, A (2011) The mythology of feedback, *Higher Education Research and Development* 30 (4), 405–419.
- Ajjawi, R and Boud, D (2018) Examining the nature and effects of feedback dialogue, *Assessment & Evaluation in Higher Education* 43 (7), 1,106–1,119.
- Bloxham, S and Campbell, L (2010) Generating dialogue in assessment feedback: Exploring the use of interactive cover sheets, *Assessment & Evaluation in Higher Education* 38 (4), 466–476.
- Boud, D and Dawson, P (2021) What feedback literate teachers do: an empirically-derived competency framework, *Assessment & Evaluation in Higher Education* 48 (2), 158–171.
- Braun, V and Clarke, V (2006) Using thematic analysis in psychology, *Qualitative Research in Psychology* 3 (2), 77–101.
- Carless, D (2006) Differing perceptions in the feedback process, *Studies in Higher Education* 31 (2), 219–233.
- Carless, D (2020) Longitudinal perspectives on students' experiences of feedback: a need for teacher–student partnerships, *Higher Education Research & Development* 39 (3), 425–438.
- Carless, D and Boud, D (2018) The development of student feedback literacy: enabling uptake of feedback, *Assessment & Evaluation in Higher Education* 43 (8), 1,315–1,325.
- Carless, D and Winstone, N (2023) Teacher feedback literacy and its interplay with student feedback literacy, *Teaching in Higher Education* 28 (1), 150–163.
- Carless, D, Salter, D, Yang, M and Lam, J (2011) Developing sustainable feedback practices, *Studies in Higher Education* 36 (4), 395–407.
- Dawson, P, Boud, D, Henderson, M, Phillips, M, Molloy, E and Ryan, T (2019) What makes for effective feedback: Staff and student perspectives, *Assessment & Evaluation in Higher Education* 44 (1), 25–36.
- Hattie, J and Timperley, H (2007) The power of feedback, *Review of Educational Research* 77 (1), 81–112.
- Hayes, H, Parchman, M L and Howard, R (2011) A logic model framework for evaluation and planning in a primary care practice-based research network (PBRN), *Journal of the American Board of Family Medicine* 24, 576–582.
- Higgins, R (2000) 'Be more critical': rethinking assessment feedback, paper presented at British Educational Research Association Conference, Cardiff University.
- Hu, G and Lei, J (2014) English-medium instruction in Chinese higher education: A case study, *Higher Education* 67, 551–567.
- Hyland, K (2019) What messages do students take from teacher feedback?, in Hyland, K and Hyland, F (Eds) *Feedback in Second Language Writing: Contexts and Issues* (Second edition), Cambridge: Cambridge University Press, 265–284.
- Hyland, F and Hyland, K (2001) Sugaring the pill: Praise and criticism in written feedback, *Journal of Second Language Writing* 10 (3), 185–212.
- Jiang, L and Yu, S (2021) Understanding changes in EFL teachers' feedback practice during COVID-19: Implications for teacher feedback literacy at a time of crisis, *The Asia-Pacific Education Researcher* 30, 509–518.
- Jiang, L, Yu, S and Zhao, Y (2019) Teacher engagement with digital multimodal composing in a Chinese tertiary EFL curriculum, *Language Teaching Research* 25 (4), 613–632.

- Jönsson, A (2013) Facilitating productive use of feedback in higher education, *Active Learning in Higher Education* 14 (1), 63–76.
- Kirkpatrick, A (2017) The languages of Higher Education in East and Southeast Asia: Will EMI lead to Englishisation?, in Fenton-Smith, B, Humphreys, P and Walkinshaw, I (Eds) *English Medium Instruction in Higher Education in Asia-Pacific: From Policy to Pedagogy*, Cham: Springer, 21–36.
- Kostka, I and Maliborska, B (2016) Using Turnitin to provide feedback on L2 writers' texts, *TESL-EJ* 20 (2), 1–22.
- Lundstrom, K and Baker, W (2009) To give is better than to receive: The benefits of peer review to the reviewer's own writing, *Journal of Second Language Writing* 18 (1), 30–43.
- Mahoney, P, Macfarlane, S and Ajjawi, R (2019) A Qualitative Synthesis of Feedback in Higher Education, *Teaching in Higher Education* 24 (2), 157–179.
- Maxwell, J A (2004) Using Qualitative Methods for Causal Explanation, *Field Methods* 16 (3), 243–264.
- Molloy, E, Boud, D and Henderson, M (2020) Developing a Learning-Centred Framework for Feedback Literacy, *Assessment & Evaluation in Higher Education* 45 (4), 527–540.
- Morse, J (1995) Editorial: The significance of saturation, *Qualitative Health Research* 5 (2), 147–149.
- Nash, R and Winstone, N E (2017) Responsibility-Sharing in the Giving and Receiving of Assessment Feedback, *Frontiers in Psychology* 8, 1519.
- Nicol, D (2010) From monologue to dialogue: improving written feedback processes in mass higher education, *Assessment & Evaluation in Higher Education* 35 (5), 501–517.
- Nicol, D, Thomson, A and Breslin, C (2014) Rethinking feedback practices in higher education: a peer review perspective, *Assessment & Evaluation in Higher Education* 39 (1), 102–122.
- O'Donovan, B (2017) How Students' Beliefs About Knowledge and Knowing Influence Their Satisfaction with Assessment and Feedback, *Higher Education* 74 (4), 617–633.
- O'Donovan, B, Rust, C and Price, M (2016) A Scholarly Approach to Solving the Feedback Dilemma in Practice, *Assessment & Evaluation in Higher Education* 41 (6), 938–949.
- Ou, W A and Gu, M M (2021) Language socialization and identity in intercultural communication: Experience of Chinese students in a transnational university in China, *International Journal of Bilingual Education and Bilingualism* 24 (3), 419–434.
- Price, M, Handley, K and Millar, J (2011) Feedback: Focusing Attention on Engagement, *Studies in Higher Education* 36 (8), 879–896.
- Robinson, S, Pope, D and Holyoak, L (2013) Can We Meet Their Expectations? Experiences and Perceptions of Feedback in First Year Undergraduate Students, *Assessment & Evaluation in Higher Education* 38 (3), 260–272.
- Saunders, M (2000) Beginning an Evaluation with RUFDATA: Theorizing a Practical Approach to Evaluation Planning, *Evaluation* 6 (1), 7–21.
- Saunders, M (2012) The use and usability of evaluation outputs: A social practice approach, *Evaluation* 18 (4), 421–436.
- Tight, M (2021) *Syntheses of Higher Education Research*, London: Bloomsbury.
- Tuckett, A G (2005) Applying thematic analysis theory to practice: a researcher's experience, *Contemporary Nurse* 19, 75–87.

- Winstone, N and Carless, D (2020) *Designing Effective Feedback Processes in Higher Education*, Oxon: Routledge.
- Winstone, N E, Nash, R, Parker, M and Rowntree, J (2017) Supporting learners' agentic engagement with feedback: a systematic review and a taxonomy of recipience processes, *Educational Psychologist* 52, 17–37.
- Winstone, N E, Nash, R A, Rowntree, J and Parker, M (2017) 'It'd be useful, but I wouldn't use it': barriers to university students' feedback seeking and recipience, *Studies in Higher Education* 42 (11), 2,026–2,041.
- Wyatt Knowlton, L and Phillips, C C (2013) *The Logic Model Guidebook: Better Strategies for Great Results*, London: Sage.
- Zhang, X and Head, K (2010) Dealing with learner reticence in the speaking class, *ELT Journal* 64 (1), 1–9.

3

Discerning the language assessment literacy of EFL teachers in Uzbekistan: A social and sociohistorical teacher cognition inquiry

David L. Chiesa

University of Georgia, USA

This chapter contributes to the discussion of language assessment literacy and addresses the context, at both the institutional and larger macro sociocultural levels, of which the language teacher is a part. It focuses on:

- Bridging the two academic fields of language teacher cognition and language assessment literacy
- Examining an under-researched population of EFL teachers
- Discerning language teachers' knowledge, beliefs, and feelings within the sociocultural contexts (i.e., micro-institutional and macro-sociocultural) in which the teaching, learning, and assessing takes place

Language assessment literacy and language teacher cognition

Assessment is an essential component of a teacher's professional practice. According to Inbar-Lourie (2013), assessment is always situated within specific institutional and policy contexts and can play a role in a language teacher's instructional practices and professional identity. Language assessment literacy (LAL) – the level of a teacher's engagement with constructing, using, and interpreting a variety of assessment procedures to make decisions about a learner's language ability (Taylor 2013) – needs to be considered in relation not only to teacher knowledge, but also to teachers' interpretive frameworks, 'which are shaped through their particular situated personal experiences, knowledge, understanding and beliefs' (Scarino 2013:322). Therefore, a promising approach to the study of LAL comes

from the research area of teacher cognition – that is, an area of inquiry that examines what language teachers know, think, believe (Borg 2006), and feel (Johnson and Golombek 2016).

The notion of the mind is an important phenomenon to analyse if one wants to understand the process of language teaching and assessing. A teacher's mental work extends beyond what can be publicly accessible through in-person, audio, or video observation (Burns, Freeman and Edwards 2015); for instance, there is a copious amount of private mental work that goes into the planning, evaluating, reacting, and deciding stages of teaching. Language teacher cognition research is diverse in its subject matter and has been conducted throughout many different L2 and foreign language education contexts. The most notable includes L2 teacher cognition about grammar teaching and grammatical terminology (Borg 1999). One subject matter area that has been underexplored in the literature – but is a critical area of language teaching – is the relationship between L2 teacher cognition and assessment, assessment practices, and language testing.

In this study, I have taken a social and sociohistorical ontological stance, and a methodological approach, to investigate how Uzbek EFL teachers' knowledge, beliefs, and feelings on LAL are being constructed, negotiated, and conceptualised over time within evolving sociocultural and sociopolitical contexts. With this study, I attempt to tap into the mental lives of one cohort of 53 participating Uzbek EFL teachers, and understand what they perceive to be valuable in terms of knowledge and skills of assessment, and what they do with assessment and why (Freeman and Johnson 2005). This investigation helps to shed light on and bring greater clarity to the concept of L2 teachers' LAL through an analysis of social, cultural, historical, and political factors.

Ontologies in L2 teacher cognition research and research questions

Research into L2 teacher cognition has changed over time, especially as related to ontological stances and methodological approaches. Four different research traditions have been identified:

- individualist (1990 onwards) – a cognitivist ontological tradition grounded in teachers' decisions and decision-making practices, thoughts, and beliefs, which are mainly analysed through quantitative means
- social (1995 onwards) – a sociocognitive paradigm in which researchers analyse the wider surroundings and how the context shapes or informs thinking, analysed mostly through qualitative means such as using diary studies

- sociohistorical (2000 onwards) – ‘thinking as a function of place and time, through interaction and negotiation with social and historical contexts’ (Burns et al 2015:589), which has been researched quite extensively through qualitative measures, including interviews and narrative inquiry
- complex/chaotic systems (2010 onwards) – a dynamic and emergent system that involves the integration of multiple interconnected elements, and has been researched qualitatively through analysis of interactions.

More specifically, I will draw on the social and sociohistorical eras in this study.

Social ontological era

The social ontological era’s conceptual unit of analysis expands from individual teachers’ epistemological views into *meanings that are situated in social contexts* (e.g., Tsui 2003). The research methodology in this era consists mostly of qualitative research methods, including *introspective methods* – ‘the process of observing and reporting one’s own thoughts, feelings, reasoning processes, and mental states’ (Nunan and Bailey 2009:285), which features a range of research methods including stimulated recalls (e.g., Yuan and Lee 2014); diary studies (e.g., Numrich 1996); and interviews (focus groups, one-on-one) with participants that can last for extended periods or on multiple occasions (e.g., Kubanyiova 2012). With the social ontological tradition, research on language teaching moved from identifying *what* teachers think, know, and believe (Borg 2003) to understanding *how* shifts in cognition happen through the process of learning to teach across professional careers, and within instructional contexts *where* those learning processes unfold (Freeman and Richards (Eds) 1996). Here, a combination of emic (insider) and etic (outsider) perspectives is used. The second era is therefore characterised as social because researchers are looking at *the conceptual changes* in thinking from a sociocognitive perspective, ‘by emphasizing how the wider surroundings or contexts, both internal to the person and external to the social setting, shapes and/or informs thinking’ (Burns et al 2015:591). In sum, teacher cognition researchers gained a conceptual shift with this era through which L2 teacher learning was viewed as socially contingent on knowledge of self, students, subject matter, curricula, and settings.

I will continue the tradition of the social research paradigm in this study by asking the following overarching research question (RQ1): *How do Uzbek EFL teachers talk about their assessment practices and justify the scores they provide for their students?* To address this broad inquiry, I will ask three related subquestions (SQs1):

Discerning the language assessment literacy of EFL teachers in Uzbekistan

- a) What do Uzbek EFL university teachers report they do when assessing their students (i.e., assessment tasks and scoring procedures)?
- b) What knowledge, beliefs, and feelings do Uzbek EFL university teachers report surrounding their assessment practices?
- c) How do Uzbek EFL university teachers' reported knowledge, beliefs, and feelings about assessment appear to shape how they assign scores to their students' work?

The broader RQ1 and SQs1a–c will examine the conceptual changes in thinking from a sociocognitive perspective by showing how the micro-institutional and macro-sociocultural contexts, both internal to the L2 teacher and external to the social setting, shape or inform thinking about assessment (Burns et al 2015:591).

Sociohistorical ontological era

A sociohistorical perspective follows the social in the early 2000s and views the research into the minds of language teachers as 'thinking as a function of place and time, through interaction and negotiation with social and historical contexts' (Burns et al 2015:589). This perspective is multifaceted and multilayered and includes time as a major variable in how research is conducted. As we move forward from individual orientations to social orientations and now to a sociohistorical orientation, we can see a more inclusive view of the language teacher mind, one which links to Borg's (2006) view of L2 teacher cognition as an integration of sources of knowledge (e.g., schooling, professional coursework, contextual factors, classroom practice). A Vygotskian sociocultural framework of mind is adopted in this ontological paradigm (Johnson 2009), which emphasises how language teaching occurs in situated social interactions between teachers' personal propensities and social practices. A qualitative research approach is mostly used with and through a co-constructed researcher–participant dialogue (e.g., Breen, Hird, Milton, Oliver and Thwaite 2001).

In addition to RQ1 and SQs1a–c, I will expand the L2 teacher cognition research agenda on assessment literacy to include the sociohistorical research generation, and will ask the overarching RQ2: *What are the macro-environmental constraints and/or affordances in Uzbekistan that could shape how EFL teachers provide meaningful assessment situations for their students?* To address RQ2, I will ask SQs2a–c:

- a) What are the macro-environmental factors Uzbek EFL university teachers report?
- b) What are Uzbek EFL university teachers' reported knowledge, beliefs, and feelings surrounding these (macro-environmental) factors?

- c) How do the reported factors appear to shape Uzbek EFL university teachers' assessment practices?

These questions will apply an eclectic view of the language teacher mind, which reflects Borg's (2006, 2012) view of L2 teacher cognition as an integration of sources of knowledge, beliefs, and feelings that relate to schooling, professional coursework, contextual factors, and classroom practices.

Methodology

Researcher positionality

A specific cultural context I have been interested in is Central Asia, a region of the world with a rich history and culture that has had a major impact on Western countries in terms of religion, politics, and economics. Within Central Asia, I have been particularly interested in Uzbekistan because of its culture, political history, and language planning and policy, which impacted the people of Uzbekistan and the professional lives of language teachers. I was assigned to work as an English Language Specialist for the U.S. Department of State at the Flying High Training Site (pseudonym) in Tashkent – the capital of Uzbekistan – to conduct a program evaluation of the newly established national in-service language teacher education program, and ultimately provide the Ministry of Higher Education with recommendations for growth. The in-service professional development curriculum consisted of the following two topics: linguistics and second language teaching. Language assessment was a light thread between these two sections, and I became curious to understand more about what teachers there know and need to know about assessment because assessment was not clearly presented in the program's curriculum. With a clearer understanding about teachers' LAL through research, I hoped to provide a stronger program evaluation but also contribute to the academic discussions of language assessment literacy and language teacher cognition, which was aside from my responsibilities as a program evaluator.

My status as an English Language Specialist who was in Uzbekistan to conduct a program evaluation was intertwined with my positionality as a researcher, and I took on both emic and etic perspectives during this process. Adopting an emic perspective was facilitated through my day-to-day interactions with the language teachers and in the discourses they used to discuss their culture and society. I took an etic perspective to the study when I analysed the data. I used pre-existing theories on L2 teacher cognition and language assessment literacy as a guiding framework to see if they applied to the Uzbekistan EFL context.

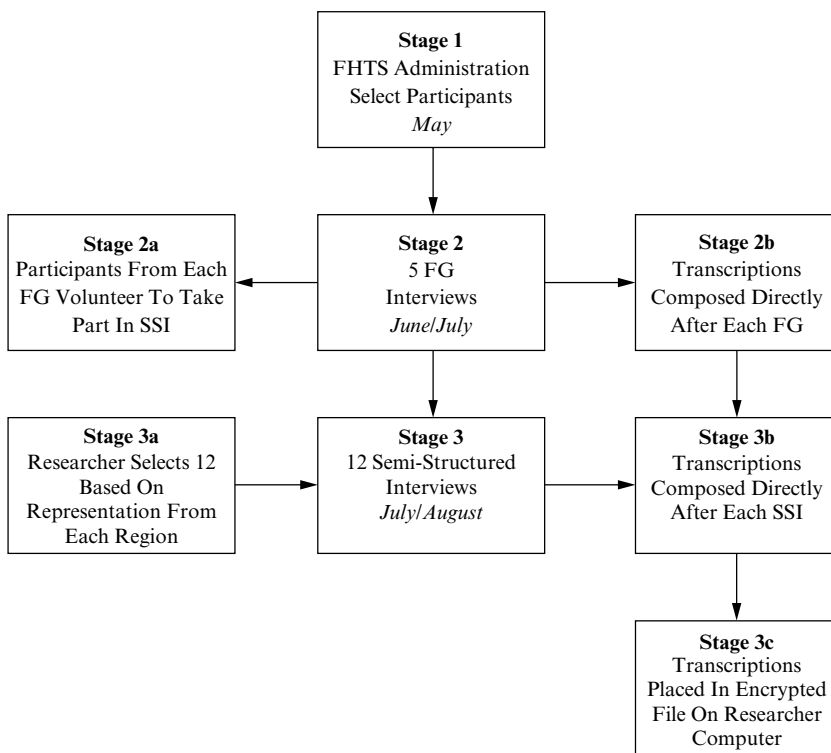
Participants

Fifty-three university English language teachers from 12 provinces and one autonomous region participated in the study at the Flying High Training Site (FHTS). All participants were provided pseudonyms for this study to protect their anonymity. Of these participants, 46 teachers were women and seven were men, which is a typical gender split in the EFL teaching context in Uzbekistan. Of these teachers, 56% taught at higher educational institutions in Tashkent while 44% worked in the regions (e.g., Andijan) and the autonomous region (Karakalpakstan). The average age of the participants was 37 and their ages ranged from 26 to 63. The average number of years of teaching experience was 11, and they ranged between one to 40 years of experience.

Data management and analysis

For both RQs I used focus groups (FG) and semi-structured interviews (SSIs) with individuals as data collection procedures. The following flowchart

Figure 1 Data collection process



shows the timeline and process for collecting audio recordings that were transcribed by the researcher.

Focus group

The FG interview was selected because it allowed me to gain a variety of perspectives and experiences from the participating Uzbek EFL teachers across the country. The 53 teachers were selected by the administration and they also made a timetable schedule, scheduled a room, and then notified me of that information. There were five different FG interviews in English, each with 10 or 11 participants; one male teacher was assigned to each group with the other participants being female.

Directly after each FG interview on the same day, I transcribed verbatim what was said in the discussion. The transcription software Dragon v. 5.0.0. for Mac was used to assist in the process. Once all transcriptions were completed, they were placed in an encrypted file on a computer with the audio recordings.

Semi-structured interviews

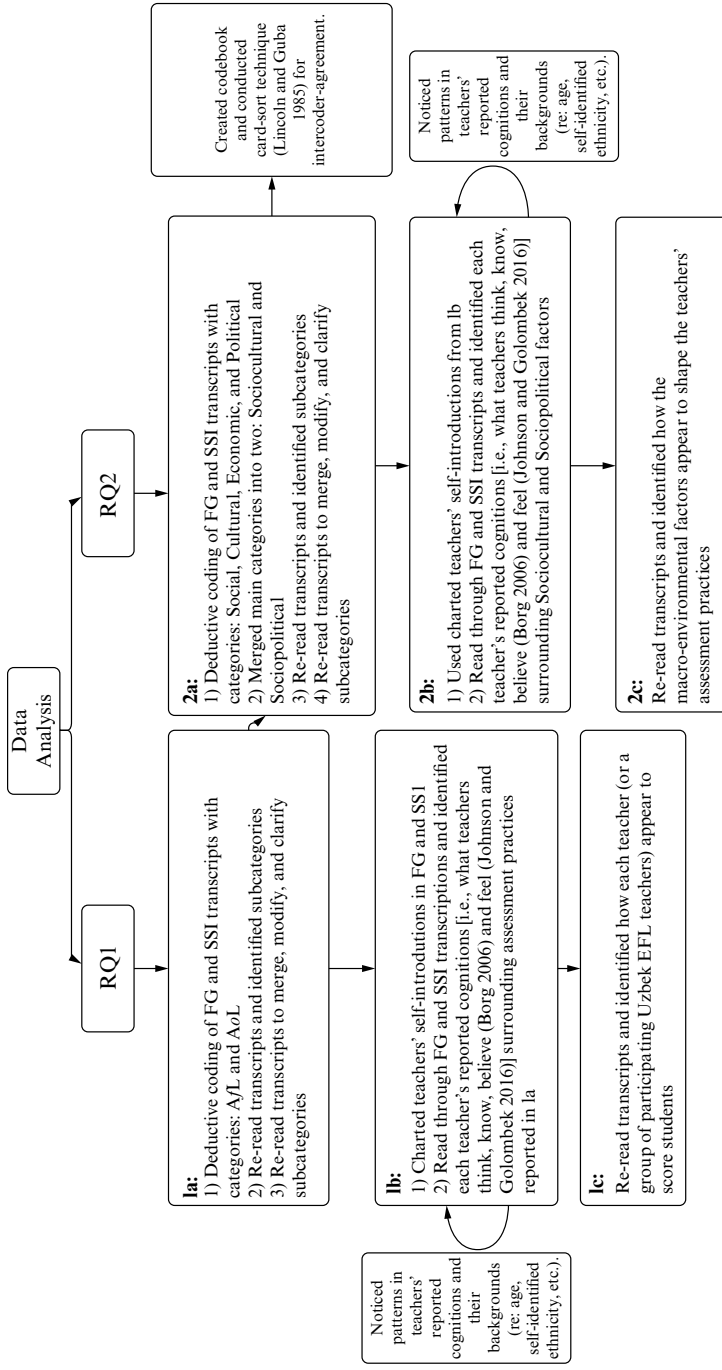
The SSIs were conducted after the FGs, because this method yielded complex data with an emphasis on subjectivity, involving the teachers' personal stories, images, and descriptions. Part of this process was the sharing of sensitive information they might not have wanted to share during the FG interview. Additionally, SSI procedures were employed because they provided a vehicle to ask clarifying questions or ask for a teacher to expand upon a point. The purpose was to uncover what the teachers do and why for language assessment. Both types of interview format provided answers to the 'why' question, or in this study, insights into the cognition behind what they did with certain assessment practices. Because of the limited time and resources available to me, I selected participants based on region, and chose 12 teachers from a group of participants who reached out to me following an invitation at the end of each FG interview. Thus, I conducted and audio-recorded 12 SSIs a week after the last FG session. Eleven participants were female and one was male.

Analysis

These focus group and interview data were subjected to close content analysis. Figure 2 below shows the data analysis process.

After combing through the FG and SSI transcripts, each line(s) or paragraph was coded deductively for SQ1a: What do Uzbek EFL university teachers report they do when assessing their students (i.e., assessment tasks and scoring procedures)? This question began the qualitative data analysis process, which started with coding for how teachers reported doing their

Figure 2 Data analysis*



*Note: *A/L* = Assessment for Learning; *AoL* = Assessment of Learning

assessment practices. Then, subcategories were created from the broad categories after the data had been read multiple times. Subsequently, to address SQs 1b and 1c – what knowledge, beliefs, and feelings do Uzbek EFL university teachers report surrounding their assessment practices; and, how do Uzbek EFL university teachers’ reported knowledge, beliefs, and feelings about assessment appear to shape how they assign scores to their students’ work – I went back through the introduction phase of each FG and SSI and charted each teacher’s self-introduction. I noted down the following: pseudonym of each teacher, what province they came from, what type of university they taught at, what type of curriculum they are a part of, and as much of their background as they revealed throughout the interviews. From these identifiers, I reread the transcripts and matched who said what at what point to discern if there were any types of patterns in the teachers’ knowledge, beliefs, and feelings, and their reported assessment practices.

After I addressed RQ1 (and its SQs) I analysed the transcripts for SQ2a: What are the macro-environmental factors Uzbek EFL university teachers report? Following the initial coding of the data (see Figure 2), I identified two core codes (i.e., Sociocultural and Sociopolitical). To make sure the core codes were consistent, I utilised the card-sort technique, initially developed by Lincoln and Guba (1985) as a quality control mechanism and conducted inter-coder agreement (.98) with an English Language Specialist (with the U.S. Department of State) in Uzbekistan who received her PhD in Applied Linguistics. Then, to answer SQs 2b and 2c – what are Uzbek EFL university teachers’ reported knowledge, beliefs, and feelings surrounding these (macro-environmental) factors; and, how do the reported factors appear to shape Uzbek EFL university teachers’ assessment practices – I used the charted categories from RQ1 and matched which teacher said what about each macro-environmental factor and when. After identifying the patterns in the teachers’ responses, I answered the overarching RQ2: What are the macro-environmental constraints and/or affordances in Uzbekistan that could shape how EFL teachers provide meaningful assessment situations for their students?

Research Question 1: Results and discussion

What do Uzbek EFL university teachers report they do when assessing their students (i.e., assessment tasks and scoring procedures)?

I classified their practices into two overarching categories: assessment pre/ during learning and assessment post-learning.

Assessment pre/during learning

The teachers described doing two things (see Table 1): (1) ongoing assessment and (2) placement/diagnostic assessments.

Table 1 Assessment pre/during learning

General category	Subcategory	Number of teachers reporting
Ongoing assessment	Presentations	30 (56%)
	Assignments	30 (56%)
	Participation	29 (55%)
Placement and diagnostic assessments		24 (45%)

Note: 53 participants.

Participants reported most of their assessment duties revolve around the evaluation of students on a day-to-day basis, which includes presentations, assignments, and participation scores. Some participants reported they find doing ongoing assessment to be a valuable, worthwhile, and necessary endeavour, while others reported that they do not understand the purpose of such assessment and have a negative belief and feeling toward it. Ajva, who has been teaching EFL for eight years at a public university, held positive beliefs/opinions: ‘continuous assessment¹ helps us teach and students learn.’ During the discussion of ongoing assessment in FG1, participants said they agreed with Ajva, suggesting they had favourable/positive beliefs toward this assessment and considered it an important teacher skill.

However, unlike the participants in FG1, some did not view ongoing assessment as favourably. Those who did not approve were primarily over the age of 50. These language teachers had previously been teachers of Russian as a foreign language while the USSR was in control of Uzbekistan’s education system. At the beginning of independence in 1991, they were told they needed to become teachers of English. Shaholo from FG4 interjected the following into the group discussion:

May I talk about my time when I was teaching in the time of the USSR? There was no continuous assessment. Because we had only two, three marks. The marks were tests. Either you came to class or you did not, and that was not [scored]. But nevertheless the level of the knowledge was much more back then. Now we have different criteria, subcriteria, what the students should know during the lesson. We divide everything

¹ There is a slight mismatch between the established use of the term ‘continuous assessment’ in our field (which is associated with formative assessment practices) and the way Uzbek EFL teachers are interpreting/applying it, which seems to mean ‘ongoing assessment practices during the teaching/learning process’ as opposed to ‘end-of-course assessment’.

to the little, little details. Too much. All the items the students should acquire, let's say. But still, we pay a lot of attention. We all the time, we tried to modify this percentage, this points for continuous assessment. If you were not trained by British Council, who made PRESETT², you do not know exactly how to score [for continuous assessment] (Shaholo).

It appears that, based on her experience during the USSR time in Uzbekistan, Shaholo believes the practice of ongoing assessment is too complicated a procedure, and that it ruins the flow of teaching and learning language. During this time, language teachers focused on disseminating knowledge to students so they could help them pass tests, because passing tests was seen as the ultimate marker of achievement. Therefore, the new teaching approach of assessing students daily appears to be a challenge for some EFL teachers unaccustomed to this practice. Shaholo's comments about the British Council suggest that she might have had a more positive/favourable outlook on ongoing assessment if her university had provided her the opportunity to be trained. Her beliefs and feelings seem to be shaped by her university context and her past/current teaching experiences within two different political climates; however, she will still provide an ongoing assessment score for her students because she is 'required to do this policy' (Shaholo).

Assessment post-learning

The second assessment category relates to post-learning, specifically, the final examination. The participating teachers' comments on this topic may be further subdivided into: 1) how they design and evaluate final exams, and 2) how they administer and score these exams (see Table 2).

Table 2 Assessment post-learning

Broad category	Sub-categories	Number of teachers reporting
Final examination	Design & evaluate	32 (60%)
	Administer & score	15 (28%)

Note: 53 teachers in total.

2 The PRESETT (Pre-Service English Teaching and Training) curriculum was created by the British Council in 2013 and has been required by the Ministry of Higher Education to be used at all Uzbek universities. The PRESETT Curriculum is designed to prepare future teachers of English to be competent in using the English language and in employing effective methods of language teaching. Courses include English language skills (e.g., listening/speaking, reading/writing, and vocabulary) and Teaching English as a Foreign Language (TEFL) (e.g., approaches to language teaching, language learning, and classroom observation).

Test design is the most common assessment practice reported by the participants. One such design practice common in Uzbekistan is the department head deciding if each teacher or a group of teachers within their English department will be charged with the design of the final exam papers for all classes. Once test papers are created, they are sent for initial review to the department head and then distributed to the students. Uzbek EFL teachers try to design their final examinations based on the curriculum requirements and/or the goals of the courses they teach.

What knowledge, beliefs, and feelings do Uzbek EFL university teachers report surrounding their assessment practices?

Some participants reported that they believe the language test design process in Uzbekistan has many flaws. First, teachers do not feel comfortable designing tests because they often lack sufficient knowledge and skills of language assessment, and they feel that taking items (particularly multiple-choice ones) from the internet is an easier and more practical way to design tests. Aisara from FG1 explained that ‘... we are not test developers. ... [T]hat’s why we try to take some activities [of] reading and writing from some books and internets. We are not going to develop it [ourselves] because we are not experts.’

Second, the participants have different opinions regarding the potential effectiveness of the multiple-choice format. Most of the participants commented on the practice of using multiple-choice items on their final examinations as these items seem relatively ‘easy to do and really easy to score’ (Abdulaziz). However, Diora from Khorezm, who is a novice EFL teacher, has been reading extensively on recently published TEFL methodology and testing. She reported that she ‘tries to use [the multiple-choice format] as less as possible on the tests ... [F]or example, I give [students a] passage from one novel, and they should guess [the writer’s] attitude, and write the answer in prose form.’ Diora believes that having students produce language (e.g., writing) would reveal to her more reliable information about their language abilities than the items of a multiple-choice test.

Unlike Diora, Almina is an older EFL teacher in Uzbekistan and has been teaching English since before the independence of 1991. Her past experiences in language test design have influenced the way she views the multiple-choice format. While participating in FG2, Almina reported that the students need to be able to do any kind of test method because the tests are grounded in the knowledge of the course. She explained that: ‘I know. I gave the information. I gave the knowledge, and according this information, the knowledge, I prepare my tests. I don’t care whether or not they can manage this, could they solve this test or no. I know that I gave them this information, and

they should know this ...’ Almina is the only person who reported that she does ‘not care’ if a test item is too difficult (or easy), because she believed the students should have learned all the information she provided for them during the course. Additionally, she does not think about the test method effect and how that could have a role in a students’ performance.

Third, some participants question the validity and/or reliability of the kinds of final tests commonly approved by the administration. Ulugbek from Djizzak, who has not had language assessment training, said he has ‘... seen mistakes in the test [papers], which were not edited properly before the exam. I have seen [problems] in those test papers which were created by other teachers.’ Ulugbek points out that the final examinations created for his university have problems. He wants to learn how to conduct item analysis, but he lacks the training and resources (e.g., materials) to acquire such knowledge. My conversation with Nodira from Karakalpakstan, who has taken language assessment courses in an MA program outside of Uzbekistan, rejects the idea that a person checks the quality of the tests at her university, and does not believe that this process happens across Uzbekistan. She explains:

I doubt they follow such procedures because the tests that I see and I am now involved in are very poorly written. I am afraid that we are not measuring accurately. Maybe we are measuring 50% accurately. We have practicality and validity issues of the tests we give our students. I doubt that they were reviewed by someone. There are so many mistakes in the questions and sometimes there are no answers ... They are not of good quality, I guess. It’s my idea, but not only mine. But many teachers think so (Nodira).

Fourth, some participants believe that the test process is unfair and does not truly measure what is being taught in all classrooms – the tests only reflect what is being taught in a few classes.

How do Uzbek EFL university teachers’ reported knowledge, beliefs, and feelings about assessment appear to shape how they assign scores to their students’ work?

All participating Uzbek EFL teachers reported cognition (i.e., knowledge, beliefs, and feelings) about language assessment, which varies from thinking that certain assessment practices are valuable and worthwhile in improving students’ language learning, to not needed and therefore not necessary. They also reported their thoughts on the test development process, which they believe has many flaws. There are five overarching factors that seem to influence the participants’ knowledge, beliefs, and feelings about language assessment practices:

Discerning the language assessment literacy of EFL teachers in Uzbekistan

- teachers' experience teaching during Soviet Era Uzbekistan (up to 1991) versus post-USSR (after 1991)
- the number of years in their career (novice, mid-career, expert) and where they have taught EFL (countryside or city), and the number of professional development courses in which they have participated
- educational experience in an MA-level TEFL program and whether it had a language assessment/testing course, and amount of time training (e.g., professional development courses) in language assessment/testing
- experience using certain item formats to test students, and using certain scoring procedures
- the amount of exposure they have had in Western/developed countries.

Although the participants had varying beliefs, they did report doing assessment practices during and after instruction. These participants, however, were unable to explain how they used the information they gleaned from the ongoing assessment (and all assessments) to inform students about ways they could improve. Many teachers reported that they scored the ongoing assessments on general impressions (i.e., rather subjectively). Teachers who did not think that ongoing assessment was helpful, however, still graded their students because they had to provide them with a score based on the universities' grading policy. In general, whether the participant teachers held positive or negative beliefs, all scored the students on the language teachers' general impressions (i.e., did not use rubrics). Thus, there seem to be discrepancies between what the participating teachers do, what they think about what they do, and how they assess students for learning.

Research Question 2: Results and discussion

What are the macro-environmental factors Uzbek EFL university teachers report?

Five macro-environmental factors have been discerned that shape Uzbek EFL teachers' assessment practice:

1. The cultural value of collectivism.
2. The presidential decrees of Karimov and Mirziyoyev.
3. Societies' perceptions of language teachers.
4. The Uzbek ethical responsibility to selves and others.
5. Influences of the British Council and the U.S. Department of State.

Regarding the first factor, the cultural value of collectivism, Madison from FG3 explained that the 'Uzbeks have such kind of value, it comes out of our mentality, cooperation ... [I]f we unite together, we will develop.' In general,

the teachers identified their culture as a collectivist society, which means they value group relationships over individual pursuits. Laden within this value is the belief reported by the participants that all students, regardless of race, class, and ethnicity (e.g., Uzbek, Kazakh, etc.) should not fail the courses they take. Instead, students who are struggling with course content, classroom assignments, and/or examinations should be provided with sufficient opportunities and support structures so they can learn the required knowledge or possess the necessary skills to pass a course. In general, the participants want the best for their students. The participants will try all that they can, including changing and improving their ways of teaching in order not to fail them, particularly, if there is a reason to better support the Uzbek culture and way of life (e.g., marriage).

Aziza described the importance of marriage to a typical Uzbek family: ‘As we know, as soon as a child is born, parents start preparation for his or her marriage by collecting *sarpa* [all necessary household items] for a girl or starting to build a house for a boy so that in 20 years everything is ready for a child to get married.’ According to Aziza, a cultural goal of an Uzbek is to continue the traditions and practices of Uzbek society, which first and foremost begins with marriage (and subsequently having children). Thus, students who get married while a course is in progress would receive more time to complete classroom assignments and tasks including class projects, papers, and/or homework assignments that can extend to a year or more beyond the last day of class.

The second macro-environmental factor that emerged are President Karimov’s (2015) and Mirziyoyev’s (2018) presidential decrees. The decrees were written to establish governmental structures to provide resources and services that influence, for the better, language teachers’ teaching, learning, and assessing practices (e.g., Exec. Order No. 24, UP-4732, 2015). The decrees constrain the education system of Uzbekistan to help provide focus and direction for what should be taught, learned, and assessed. The teachers believe the decrees have impacted their lives for the better with the creation and establishment of new educational standards that the EFL teachers must learn (i.e., National Standards, Common European Framework of Reference for Languages), and be able to apply to their teaching contexts. Most participants also noted that they appreciate the direction the country is moving toward a more Western-style education system (i.e., standards-based).

The third macro-environmental factor that emerged is societal perceptions of language teachers. An English language teacher in Uzbekistan is perceived as someone who is a master of the language and culture, and who has access to a worldwide network of academic, political, and business knowledge. Svetlana reported that her mother believed that because of a teacher’s power, teaching ‘is a sacred profession.’ Language educators are first and foremost

teachers, who hold a certain level of power and control of the classroom, the content and information provided to the students, and how they test students.

The fourth macro-environmental factor is the ethics (moral code) of Uzbek EFL teachers. A teacher's ethics serve to shape meaningful language assessment practices. During the focus groups and individual interviews, most teachers reported that they believed their ethics (morality) is a result of parental influence. Family members taught them how to treat others and how they should expect to be treated in turn. The underlying lesson often repeated by the participating teachers is that others should be treated fairly and without bias, regardless of whatever the perceived outcomes of their actions might be. A common occurrence reported by all participants in the semi-structured interviews was that all Uzbek EFL teachers would be faced with an ethical dilemma – accepting or rejecting money in exchange for a higher score. For example, Diora, like most of the Uzbek language teachers during the SSIs, became upset at the thought of bribery. I asked her how she felt when a student tries to offer money in exchange for a higher grade:

I'm usually very rude and I cut that off from the first beginning. I'll be like, "What am I doing here? Am I in business or am I trading something, or am I teaching you? Am I training you to do this in the future and is this going to be the same when you teach? How are you going to feel when you are treated like that when you are a teacher?" And they get the message.

Diora's decision to reject her students' bribe reveals an ethical decision, and suggests her more principled beliefs about what it means to be a responsible teacher. Diora believed that if she was able to help her students recognise the problem of bribery, then she would also be helping society move away from corruption. All teachers reported the current state of corruption in the Uzbek education system; however, they also reported that far fewer teachers currently solicit or accept bribes, and with each day moving forward, more teachers are opting out of the bribery system and only assign the grades students have earned.

The fifth macro-environmental factor that could shape assessment conditions comes from the knowledge/skill base of outside/foreign entities. Foreign groups from the British Council and the U.S. Department of State provide multiple resources to support Uzbek EFL teachers' LAL. The participating teachers report that the knowledge base learned from teacher educators from the British Council and the U.S. Department of State constitute outside influences that have not been constraining the assessment practices of Uzbek EFL teachers. Rather, they have been providing teachers with opportunities to explore different assessment practices that are meaningful and relevant to local contexts of language teaching.

What are Uzbek EFL university teachers' reported knowledge, beliefs, and feelings surrounding these (macro-environmental) factors and how do they seem to shape assessment practices?

Through analysis I identified a cultural and historical pattern in their thoughts, which was verbally expressed in their reported spoken language. The conceptual anchor for expressing their knowledge, beliefs, and feelings about sociocultural and sociopolitical aspects revolves around their personal relationship to Russia and/or Russia's influence in Uzbekistan's education system. For example, Mohira in her individual interview reported that bribery did not exist in the past because Uzbeks were part of a 'more Islamic society before the USSR, and Islam does not allow bribery'; however, it was the USSR, which 'introduced a lot of corruption within all spheres of Uzbekistan, including [its] educational system, so, beginning from entrance exams, assessing, [to] assigning [grades]. Bribery has been the lead factor, so this was a bad influence. This all comes from the USSR regime' (Mohira). Another example comes from the interview with Nozliya from Syrdarya who teaches in the PRESETT curriculum and who is Russian-born and educated. She views the axiom that 'students should not fail' as antiquated because 'this was not practiced in the Russian education system' (Nozliya). During our one-on-one interview, we had a conversation about her experience failing students and the cultural backlash she received from her Uzbek colleagues. Nozliya explains that when a student fails 'it is not seen as a good thing. It's not a good thing for both student and teacher. Teachers are not liked or welcomed when they fail students. I was not liked.'

In this study, participants made a direct or indirect reference to the USSR. Their reference was reported in a positive and/or negative way, and thus, participants often compared life experiences either (1) while Russia was a significant force in Uzbekistan's educational system, (2) where the country is now in relationship to Russia, or (3) where they believe the country of Uzbekistan will be in the future and its role on the world stage. There are five overarching patterns identified in the teachers' reported beliefs and feelings toward the sociocultural/sociopolitical macro-environmental facets that seem to influence their assessment practices:

- Russian-born/educated Uzbek EFL participants often did not comply with the culturally accepted rule of not failing students and were often ostracised from their Uzbek-ethnic born peers if they did not pass underperforming students.
- Many of the participants consider the relationship between themselves and their students as a parent-child relationship and view the classroom

as an extension of Uzbek society. Thus, they often blame themselves when students underperform.

- Ethics (one's moral code) is directly connected to how a person was raised. All participating Uzbek EFL teachers reported that being fair to each person in society was important. However, some participants identified fairness in education differently and *all* SSI participants commented on corruption in the education system. The participating Uzbek EFL teachers also commented that the cultural impact of corruption is waning slowly and will eventually be non-existent.
- The presidential decrees of Karimov and Mirziyoyev are moving the education system of Uzbekistan away from Russian-style education, which directly influences the type of specialists who are coming to Uzbekistan to support curriculum creation, development, and integration.
- All participants report that they appreciate the knowledge coming from British Council and the U.S. Department of State, and would like to know more about contemporary principles of language assessment and testing from these organisations.

As identified above, these five overarching patterns correspond to Borg's (2003) conceptualisation of language teacher cognition. As explained above, Borg's (2003) use of the term *cognition* reflects an integration of sources of knowledge, which includes schooling, professional coursework, contextual factors, and classroom practice. The categories listed above correspond to the category of contextual factors that seem to influence L2 teacher knowledge, beliefs, and feelings.

Bridging L2 teacher cognition and language assessment literacy

In conducting the study, I aspired to bridge the relationship between L2 teacher cognition research and LAL research. I wanted to better understand what Uzbek EFL teachers perceive to be valuable in terms of knowledge and skills of assessment, what they do (with assessment), and why. The study furthered the conceptual clarity of L2 teachers' LAL with the analysis of social, cultural, historical, and political factors. Based on the qualitative results, I suggest a new definition for LAL for classroom language teachers that bridges the two fields. LAL for language teachers is:

The capacity to be aware of one's knowledge, beliefs, and feelings toward assessment/testing practices, and to have the ability to act upon these facets in order to create relevant and meaningful assessment experiences

for learners that are situated within specific local sociocultural and sociopolitical contexts.

In other words, being language assessment literate is to have a connection and/or to build a connection between what teachers do, what they think about what they do, and how they assess students in a relevant, meaningful, and ethical way. Additionally, to be assessment literate, a language teacher has to be able to make necessary changes in their assessment practices so that language learners may benefit from such changes.

References

- Borg, S (1999) Studying teacher cognition in second language grammar teaching, *System* 27 (1), 19–31.
- Borg, S (2003) Teacher cognition in grammar teaching: a literature review, *Language Awareness* 12 (2), 96–108.
- Borg, S (2006) *Teacher cognition and Language Education: Research and Practice*, London: Continuum.
- Borg, S (2012) Current approaches to language teacher cognition research: A methodological analysis, in Barnard, R and Burns, A (Eds) *Researching Language Teacher Cognition and Practice: International Case Studies*, Bristol: Multilingual Matters, 11–29.
- Breen, M P, Hird, B, Milton, M, Oliver, R and Thwaite, A (2001) Making sense of language teaching: Teachers' principle and classroom practices, *Applied Linguistics* 22, 470–501.
- Burns, A, Freeman, D, and Edwards, E (2015) Theorizing and studying the language teaching mind: Mapping research on language teacher cognition, *The Modern Language Journal* 99 (3), 585–601.
- Exec. Order No. 24 – UP-4732, Collection of Laws of the Republic of Uzbekistan (12 June 2015).
- Freeman, D and Johnson, K E (2005) Towards linking teacher knowledge and student learning, in Tedick, D J (Ed) *Language Teacher Education: International Perspectives on Research and Practice*, Mahwah: Lawrence Erlbaum, 73–95.
- Freeman, D and Richards, J C (Eds) (1996) *Teacher Learning in Language Teaching*, Cambridge: Cambridge University Press.
- Inbar-Lourie, O (2013) Language assessment literacy, in Chapelle, C A (Ed) *The Encyclopedia of Applied Linguistics*, 2,923–2,931.
- Johnson, K E (2009) *Second Language Teacher Education: A Sociocultural Perspective*, New York/London: Routledge.
- Johnson, K E and Golombek, P R (2016) *Mindful L2 Teacher Education: A Sociocultural Perspective on Cultivating Teachers' Professional Development*, New York: Routledge.
- Kubanyiova, M (2012) *Teacher Development in Action: Understanding Language Teachers' Conceptual Change*, Basingstoke: Palgrave Macmillan.
- Lincoln, M H and Guba, E G (1985) *Naturalistic Inquiry*, Newbury Park: Sage.
- Numrich, C (1996) On becoming a language teacher: Insights from diary studies, *TESOL Quarterly* 30, 131–153.

Discerning the language assessment literacy of EFL teachers in Uzbekistan

- Nunan, D and Bailey, K M (2009) *Exploring Second Language Classroom Research: A Comprehensive Guide*, Boston: Heinle, Cengage Learning.
- 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.
- Taylor, L (2013) Communicating the theory, practice, and principles of language testing to test stakeholders: Some reflections, *Language Testing* 30 (3), 403–412.
- Tsui, A B M (2003) *Understanding Expertise in Teaching: Case Studies of Second Language Teachers*, Cambridge: Cambridge University Press.
- Yuan, R and Lee, I (2014) Pre-service teachers' changing beliefs in the teaching Practicum: Three cases in an EFL context, *System* 44, 1–12.

4 How do teacher educators learn about and teach language assessment?

Salomé Villa Larenas

Universidad Alberto Hurtado, Chile

Tineke Brunfaut

Lancaster University, UK

This chapter explores and discusses the development of language assessment competence within the teacher education context. By means of a study on teacher educators in Chile, it sheds light on:

- How teacher educators learn about language assessment
- How teacher educators teach language assessment to the next generation of language teachers
- The concept of teacher learning, with reference to language assessment, as seen through the lens of the sociocultural theory of learning
- The importance of communities of practice in learning about language assessment
- The role of modelling and reflective practice in teacher education to develop language assessment competence

Este capítulo explora y analiza el desarrollo de la competencia en evaluación de idiomas (*language assessment literacy*) en el contexto de la formación de profesores de inglés. A través de un estudio sobre formadores docentes en Chile, el capítulo arroja luz sobre:

- Cómo los formadores docentes aprenden sobre evaluación de idiomas
- Cómo los formadores docentes enseñan evaluación de idiomas a la próxima generación de profesores de inglés
- El concepto de aprendizaje docente, con relación a la evaluación de idiomas, visto a través desde la perspectiva de la teoría del aprendizaje sociocultural

- La importancia de las comunidades de práctica en el aprendizaje de la evaluación de idiomas
- El rol del modelaje y la práctica reflexiva en la formación docente para el desarrollo de la competencia en evaluación de idiomas

Language assessment literacy and teacher education

Language assessment literacy (LAL) is essential for language teachers' practice and should 'form an integral part of [their] professional development' (Harding and Kremmel 2016:415). However, the LAL literature provides evidence of language teachers' generalised feelings of underpreparedness for their language assessment practices (e.g. Berry, Sheehan and Munro 2019, Gardner and Rea-Dickins 2001, Tsigari and Vogt 2017). One reason which could explain these feelings is the generally insufficient language assessment training offered in teacher education – an issue which has been reported across the world. For example, Lam (2015) found that teacher education programmes in Hong Kong offered insufficient support to promulgate LAL and included limited language assessment training. He explained that 'not every [teacher education institution] offers mandatory language assessment-focused [sic] courses to train prospective teachers, not to mention a lack of government initiatives to enforce the requirement of LAL as part of regular teacher qualifications' (Lam 2015:190). With reference to the Colombian context, López and Bernal (2009) also found that language testing training was scant; it was offered in only 14 of the 34 teacher education programmes, and only 32 of the 82 in-service English teachers in this study reported having received training in language assessment. Similarly, a limited training offering has been reported for the European context. For example, Vogt and Tsigari (2014) conducted a study across seven European countries on language teachers' perceived LAL levels and language assessment needs. They found that although 68.3% of the 878 surveyed language teachers had received some general training in language assessment, many reported little training, or no training at all in more specific areas (e.g. on preparing classroom tests (66.2%), giving grades (45.1%), placing students into courses (50.5%), or awarding final certificates (58.2%)).

Many recommendations for LAL development put forward in the literature rest on the idea of enhancing teacher education programmes to improve the assessment knowledge and practices of teachers. For instance, Vogt and Tsigari (2014) argued that improving teacher training programmes in the area of language assessment can ensure the quality of teachers' language assessment practices, both in pre-service and in-service contexts. Levy-Vered

and Alhija (2015:393), in a study on the effects of teacher training on student teachers' conceptions of assessment and their levels of assessment literacy, found that, 'provided that training is of better quality (acquired knowledge, practising, and modelling), the assessment literacy of beginning teachers is higher, and the conceptions are more positive'. The authors consequently argued that the development of assessment literacy can be optimised through teacher education programmes by: '(a) providing professional knowledge and promoting skills related to student assessment ... (b) creating opportunities for practical assessment experiences ...; and (c) modelling quality-controlled and appropriate assessment practices by the teacher educators' (Levy-Vered and Alhija 2015:382).

Some valuable efforts have in fact been made to map out the type of content that is crucial for developing LAL, i.e. *what* needs to be covered in training. For example, Brindley (2001) theoretically outlined essential LAL components for professional development programmes: definitions and descriptions of language proficiency, construction and evaluation of language tests, assessment in the language curriculum, assessment practices, and the social context of assessment. Davies (2008) outlined changes and trends in the content covered by language testing textbooks from 1960 until the first part of the 2000s.

However, to date, little is known about *how* that content – in other words how language assessment – is taught and learned within the teacher education context (see also Brown and Bailey 2008 for similar concerns). In this regard, Graham (2005) stressed the role of the teacher educator in the development of student teachers' assessment competences. She investigated the influence of mentors' classroom assessment theories and practices on those of their student teachers. By documenting the changes in classroom-based assessment practices of two cohorts of student teachers, Graham (2005:614) found that '[a]lthough teacher candidates identified many different influences on their thinking about assessment, in many cases they reported their mentor teachers as the single most powerful influence'.

The role of teacher educators then might be pivotal to sound LAL development in language teachers. However, limited research has focused on the teacher educator and their impact in the area of (language) assessment competence development. As Murray (2016:35) asserted: 'teacher educators ... remain an under-researched, poorly understood, and ill-defined occupational group.' Consequently, little is known about how and in what ways teacher educators train future language teachers in language assessment matters. In fact, little is also known about teacher educators' own LAL and their competence development in this area.

Our interest, therefore, lay in exploring not only what but also how teacher educators teach language assessment to their student teachers, and how they have learned about language assessment themselves in the first place.

An exploration of how teacher educators learn about and teach language assessment: An example from Chile

The above interest formed part of a larger study we conducted on the LAL of teacher educators (see Villa Larenas 2020, Villa Larenas and Brunfaut 2022). In the present chapter, we exclusively focus on the part of our research that investigated how teacher educators learned about language assessment themselves, and what and how they now teach about it in their courses. The study was set in the Chilean context, and more specifically that of English as a Foreign Language teacher education (henceforth, EFLTE). For contextual and practical reasons, the focus was primarily on formal, marked assessments.

In Chile, EFLTE is commonly offered as a five-year undergraduate degree programme. By the end of the programme, graduates are expected to have the competences to teach English in Grades 5–12 of the Chilean school system and to have C1 level English proficiency on the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001). Consequently, EFLTE programmes offer training in both language pedagogy and English as a Foreign Language proficiency (Chile's official language is Spanish). In theory, no specific qualification is required to become an EFL teacher educator in these programmes, although in practice, due to the accreditation criteria for teacher education programmes in Chile, EFL teacher educators need to hold a Master's degree in ELT or a linguistics-related area and have experience teaching in Chilean schools.

Twenty teacher educators from six EFLTE programmes around the country took part in our research. They were qualified language teachers with TESOL Master's degrees (except for two who held general linguistics Master's degrees). In the Chilean context, two types of teacher educator roles can be identified that are relevant to LAL training; we label these here as *testing* teacher educators and *language* teacher educators. *Testing* teacher educators are those professionals who teach language assessment to future teachers. In Chilean EFLTE programmes, this training is either through language testing courses (52% of Chilean programmes; Villa Larenas 2020) or, in the absence of such courses, through teaching methodology courses which should include language assessment as part of their learning content. When differentiation between these two types of courses is relevant for this chapter, we will refer to those involved as language-testing-course teacher educators (three participants) versus methodology-course teacher educators (three participants). *Language* teacher educators, on the other hand, are those professionals who teach English language acquisition courses to develop student teachers' English proficiency. EFL is usually taught throughout all five years of teacher training in Chile, and those teaching these courses spend

the most time with student teachers during their training – both to develop student teachers' English language proficiency and to help student teachers develop in general matters concerning language learning, teaching, and assessment (which the student teachers can observe from the language teacher educators' teaching and assessment practices on the EFLTE programme). Fourteen language teacher educators were involved in our study.

In the research reported in this chapter, we were interested in exploring the teaching and learning of language assessment by both groups of teacher educators. To this end, we conducted individual interviews with the 20 participants (testing teacher educators and language teacher educators). The interviews comprised a semi-structured and an unstructured phase, and were audio-recorded and then transcribed. For the former phase, we developed an interview guide with questions regarding teacher educators' professional background and their role as teacher educators, their previous training in (language) assessment, how language assessment was approached/taught in their institutions, their own language assessment practices, and what they taught about language assessment in their courses and how they did that. For the latter phase, we asked the teacher educators to bring along some of their course materials (e.g. syllabuses, test specifications, test task instructions, rating scales) and talk to us about the materials' purposes, development, and uses. Through these materials, we hoped to gain a better understanding of the assessment contents and instruments chosen or developed by the teacher educators.

We thematically coded the interview transcripts and course materials' content, using a primarily data-driven coding scheme. The scheme was developed over three cycles of coding, guided by central themes such as those related to learning about language assessment and teaching of language assessment. The final coding scheme, for the part of the research reported here, comprised three high-level codes with multiple subcodes each. The scheme can be found in the Appendix. Twenty-five percent of the data was double-coded, achieving .80 inter-coder agreement (classified as 'good' by Mackey and Gass 2016).

LAL development in Chilean teacher education: Findings

How teacher educators learn about language assessment

We found that the teacher educators had mainly learned about language assessment by informal means, rather than through formal language testing courses. None of the teacher educators had received formal language assessment training. Nevertheless, 17 (of the 20 participants) had taken general educational assessment courses during their own teacher education,

How do teacher educators learn about and teach language assessment?

but they felt that such training did not meet the needs of their language teaching and teacher education practices, as it was disconnected from their discipline. For example, Teacher Educator 6 (henceforth TE6) said:

I think that we need more things about “language” evaluation ... because it was general, the [course] that I told you about. I remember we had to create an evaluation instrument, using, for example, the different types of questions that we could make, or create like multiple-choice and those kinds of things. But it wasn’t about English because the students were from different programmes. So, it wasn’t very useful for us because I think that what we needed is how to assess the different linguistic abilities because it’s important and we’re not taught that.

Unsurprisingly, the teacher educators signalled generalised feelings of insecurity in their language testing practices. For example, when asked about how they felt about their language assessment practices, TE5 said: ‘I think that I don’t know much. I feel like, in terms of theory, I know very little.’ Or, while being somewhat more confident, TE4 reflected: ‘not very confident ... well, I don’t feel “that” insecure, either. But of course, you always need more training.’

Consequently, the teacher educators resorted to other ways of learning about assessment. For example, seven reported doing self-study (e.g. by reading publications or searching for assessment practices online). Importantly, all teacher educators emphasised having learned about language assessment on the job. They considered this to have been particularly fruitful when they started their careers, to become acquainted with their programme’s language assessment culture. For example, TE3 remembered learning from a more experienced colleague:

I started here in 2012, and for the first year and a half I was coordinating with one other teacher. Basically, I was very young, it was my first teaching job. I’d been brought in to replace someone at very short notice. So, they put me on the course with someone who got so much experience. And she showed me the ropes, and basically walked me through the syllabus, the tests, the tasks and the rubrics.

More experienced teacher educators similarly valued opportunities to collaborate with other colleagues. For example, TE8, who had 15 years of experience as a teacher educator, said: ‘We all teachers contribute ideas and analyse if it is a good criterion or descriptors ... we have weekly meetings ... and we learn from each other.’ In four of the six programmes involved in this study, these collaborations were purposefully encouraged through the official formation of teams; co-teaching and teamwork were promoted. In the other two programmes, collaboration was not explicitly fostered, but three teacher educators mentioned nevertheless informally

working together with colleagues to develop curricula and conduct assessments. They felt this was helpful as they ‘did not feel so alone’ (TE1) in their assessment practices. TE8, for example, elaborated: ‘Working with a peer is very good. There are some colleagues that are very good at giving feedback ... In fact, this course is the fourth time I give it. And on previous occasions, for the [speaking assessment], I’d always asked other teachers [to assess with me].’

In both the formal and informal team settings, teacher educators talked about these collaborations as rich spaces for meaningful learning about language assessment, for example, through sharing ideas for test/task development, co-constructing tasks, and discussing marking. For instance, TE4 said:

[Learning how to assess languages] has been a self-taught process mostly, working with colleagues. They teach you a lot. So, I’ve learned with them mostly. I’d say that the experience that I had at [workplace] was quite significant for me in this regard. There, I had to work with teams doing the lesson planning process, the creation of the assessment instruments, and marking things, giving feedback. So, I’d say that most of my training in assessment ... was a collaborative process that took place there.

A further advantage of working in teams, as expressed by the teacher educators, was that it offers opportunities for learning about language assessment through reflection. They reported that reflection on their assessment practices mainly occurred in conversations within their teams, through interaction and exchange of ideas for their decision-making on possible flaws in assessment procedures and potential solutions. TE6, for example, described how coordinating with another colleague encouraged them to ‘think about assessments’, such as about better and innovative tasks, or ways to improve and to better assess what was taught:

With [other TE’s name], we are always coordinated ... For example, most of the students failed in this item ... So, we have to think, if they are failing, what the problem is. So, we discuss a lot about why ... We decided not to include this item in the next test, and we changed it by another one to see what would happen with the results. Sometimes we have better results when we have changed the items where they have failed.

Nonetheless, teacher educators also mentioned occasional instances of individual reflection on their language assessment practices, which mainly happened when the assessment results were not the desired ones. For example, TE5 said: ‘I usually reflect when I see the results, “this was useful”, or “this was too hard” or “maybe, I should have less exercises of this type”.’ Overall, however, the social interaction within their working teams was

How do teacher educators learn about and teach language assessment?

thought to offer greater opportunities for learning through reflection on one's own language assessment practices.

What teacher educators teach about language assessment

To gain insights into teacher educators' training of the next generation of language teachers in the area of language assessment, we investigated what they teach about language assessment in their courses and how they do that. Here, we focus on *what* they teach; the next section describes *how*.

To establish what is being taught about language assessment, we looked into the syllabuses and assessment materials of language testing courses and teaching methodology courses, as this is where language assessment subject knowledge is officially covered in the EFLTE programmes (not in the English acquisition courses). We analysed the subject content of these courses, how the course content was assessed, and how many hours of instruction (sessions) were dedicated to language assessment issues in these courses. Table 1 provides an overview of our findings.

As Table 1 demonstrates, the language assessment training differed greatly between the teaching methodology and the language testing courses. Firstly, teaching methodology courses dedicated little time to the teaching of (language) assessment (two to four sessions per semester). Conversely, student teachers in programmes with language testing courses enjoyed a complete semester of learning about language assessment. Secondly, there were clear differences between the content matter taught. Teaching methodology courses covered a narrow range of topics, mainly focusing on formative vs. summative assessment and test design, with a large emphasis on assessment for learning. In contrast, language testing courses went into more depth regarding matters of assessing languages, with emphasis on principles in language assessment, language constructs, and the construction and evaluation of language assessment instruments.

When looking at what these courses assessed about their subject matter and how, we again observed differences between the two types of courses. In the teaching methodology courses, the teacher educators assessed their student teachers by means of teaching-knowledge tests or quizzes, projects, and micro-teaching tasks which were connected to the student teachers' practicum settings. In other words, these course assessments evaluated teaching knowledge and/or practices, *not* student teachers' language assessment knowledge and practices. In contrast, in the language testing courses, the teacher educators purposely developed course assessments that focused on the language assessment contents and practices they taught in their course. This included tests to assess language assessment knowledge and its application, practical test development tasks, and written reports on critical analyses of language assessment instruments. For example, TE13

Table 1 Language assessment training content in EFLTE programmes

Case	Type of course	Sessions per semester	Content	Course assessment tasks
1	Teaching methodology	4	<ul style="list-style-type: none"> - assessment vs evaluation - formative vs summative - evaluation types - test construction 	<ul style="list-style-type: none"> - oral presentation - teaching content test - microteaching
2	Teaching methodology	4	<ul style="list-style-type: none"> - evaluation and assessment difference - assessment types: formative and summative - test design - assessment for special education needs 	<ul style="list-style-type: none"> - teaching content quiz - microteaching - project
3	Teaching methodology	2	<ul style="list-style-type: none"> - assessment task analysis - assessment for learning 	<ul style="list-style-type: none"> - self-assessment - school project - micro-teaching
4	Language testing	16 approx.*	<ul style="list-style-type: none"> - key principles in language assessment - strategies for language assessment criteria construction - feedback - different forms of assessment: diagnostic, formative/summative - language instruments design 	<ul style="list-style-type: none"> - rubric development project - test development projects (2) - final exam
5	Language testing	16 approx.*	<ul style="list-style-type: none"> - evaluation, assessment and testing - assessment principles - types of language tests - formative and summative - connection teaching and assessment - instruments and procedures - language construct types - item types - young learners' oral performance - assessment and critical pedagogy - designing language assessment instruments - language assessment criteria – rubrics - feedback - analysis of test results 	<ul style="list-style-type: none"> - language assessment content tests (2) - critical review of tests project - self-evaluation - test design final exam

Table 1 (continued)

Case	Type of course	Sessions per semester	Content	Course assessment tasks
6	Language testing	16 approx.*	<ul style="list-style-type: none"> - language assessment: principles - assessment of receptive and productive skills - assessment for and of learning - feedback - language assessment analysis and construction 	<ul style="list-style-type: none"> - language assessment content test - research paper - instrument design - final exam

* *The number of sessions varies depending on factors such as bank holidays.*

explained: ‘Assignment 3 was a group analysis in which they had to bring in assessment tools they had seen in their practicum; bring it to class and analyse it together with a group.’

In other words, considerable differences exist between what teacher educators teach about language assessment depending on whether they (and their university) offer a teaching methodology or language testing course – with the latter covering a wider range of language assessment content and practices in more intensive and in-depth ways.

How teacher educators teach language assessment

The scope of what constitutes EFLTE has extended over the last two decades from a focus on teaching-knowledge and skills to encompassing the development of teacher identity, the examination of language teaching processes, and the inclusion of relevant sociocultural considerations (Freeman 2009). Accordingly, different teacher education models have been proposed, e.g. the Craft Model, the Applied-science Model, and the Reflective Model (Wallace 1991). While there is no agreement on a ‘perfect’ model, two strategies – which are central to the present study – consistently feature in the teacher education literature, i.e. modelling and the encouragement of reflective practice. Through the interviews and the teacher educators’ course materials, we explored these two strategies in relation to *how* teacher educators teach language assessment, and report our findings below.

Teacher educators’ modelling of language assessment practices

Modelling is understood as the ‘practice of intentionally displaying certain teaching behaviour with the aim of promoting student teachers’ professional learning’ (Lunenburg, Korthagen and Swennen 2007:589). It is considered a ‘desirable professional competency’ (Loughran and Berry 2005:193) and to

play ‘a central role’ in teacher education (Singh and Richards 2009:204). It is a common strategy (Woodward 1991) which has been promoted in teacher education models such as the Craft Model (Wallace 1991), Apprenticeship of Observation (Lortie 1975), and Mentoring (Malderez 2009).

In the interviews, all teacher educators acknowledged that modelling is an important strategy in their practices. For example, TE5 said: ‘we create activities together [with student teachers] to learn how to create and design activities or handouts or exercises. And then we discuss them.’ However, most teacher educators (14) thought they mainly modelled teacher skills and teaching methodology, not assessment practices. TE7 said: ‘[the modelling of assessment practices] is not very explicit. We’re putting a lot of emphasis basically on the teaching part. But not in the way we assess ... it’s not explicit.’ In fact, three *language* teacher educators stated that they did not think they were modelling any assessment practices to their student teachers.

Nevertheless, the data showed that *language* teacher educators in fact unintentionally or unconsciously modelled language assessment practices in their EFL classes. This modelling usually happened when they involved student teachers in the co-construction of their language assessment materials, such as rubrics, instructions for tasks, and test items. For example, TE8 unconsciously modelled rating scale development through eliciting student feedback during scale construction for the EFL course’s assessment purposes: ‘I developed [the rubrics]. And also, students participated. I have asked the students if they understand them, asked them if they understand what they are referring to. They give me examples.’

Intentional modelling of language assessment practices was only reported by the *testing* teacher educators. For example, TE13 mentioned modelling their statistical analysis of the course’s test results for the student teachers:

I don’t see how I could ask them to do [a statistical analysis] in the final project and not do it myself. Part of my reason for doing it at all was to be able to share it with them ... so they could see what I had done, and they could see the math. So, it was part of a learning process for all of us.

Teacher educators’ encouragement of reflection about language assessment practices

Reflection is ‘the systematic and deliberate thinking back over one’s actions’ (Russell and Munby 1992:3). The reflective practice movement has become central to teacher education over the years; in fact, the encouragement of reflection on what is being taught (a metacognitive skill) is now considered an essential part of the teacher educator’s role (e.g. Burns and Richards 2009). Encouragement of reflection is thereby understood as the practice of

How do teacher educators learn about and teach language assessment?

opening spaces for reflection on what is being taught, i.e. ‘making the teacher educator’s reflection on practice explicit for the student teachers so that they can observe the thinking that underpins the learning they experience’ (Loughran 1995:431).

In our study, all interviewees argued that opening spaces for reflection about their own practices as teacher educators encourages student teacher learning and constitutes an important part of their roles as teacher educators. However, similar to the results about modelling, differences could be observed between teacher educators depending on the type of courses they teach. Namely, *language* teacher educators reported that the reflection instances they create in their classes focus on language teaching practices, teaching methodologies, and teacher skills – not language assessment practices. TE19 explained: ‘We work a lot with awareness, not only in methodology but also in the EFL [courses] and practicum. We are aiming at the students becoming aware, not only of the content ... becoming aware of how you learn the language, how you learn about the language, and how you teach.’ But when asked about reflection activities around language assessment practices (e.g., Do you talk about the tests with your students later in the class? About the purpose of the test, the design of the items, why this test is like this?), TE4 replied: ‘No, we don’t really do that.’

Within the group of *testing* teacher educators, *methodology-course* teacher educators reported offering similar reflection instances to the language teacher educators, i.e. focusing on teaching practices and skills. Only *language-testing-course* teacher educators reported deliberately encouraging reflection on language assessment practices in their student teachers. They design assessment tasks and activities in their courses which specifically elicit reflection on the process of language assessment and the design of language assessment materials. For example, TE14 described creating instances for reflection in both the language testing course and EFL course this teacher educator taught:

As a teacher trainer, I think that this [reflective phase] is a meta-cognitive one in which one tries to explain to students the logic behind each of the decisions we are making ... I think [it] is important because it helps students make connections with what they see even in teaching methodology ... This whole phase of reflection is what we achieved in the [language testing] course and at a level that they themselves create a rubric, or that they themselves analyse the assessments that other people make, since it’s like the highest level of reflection. Also, in my practice as an English teacher for the English courses, I regularly make them reflect on the assessment activities that are done.

Broader perspectives on LAL development in teacher education: Discussion

Teacher educators' *own* LAL development through the lens of sociocultural theory of learning

Our findings indicate that one of the richest spaces for teacher educators' learning about language assessment is through collaboration with colleagues. Such learning through teamwork resonates with sociocultural theories of learning. These maintain that learning occurs during social interactions between individuals (Röehler and Cantlon 1997, Vygotsky 1978), whereby 'learning is assumed to be social and situated; often occurring in informal contexts such as communities through interaction, communication, taking part, and gaining access to different contexts' (Patton and Parker 2017:352). Essentially, in the teacher education context, collaboration between colleagues can be considered a form of professional learning (e.g. Johnson and Golombek 2011). In fact, in settings such as the Chilean EFLTE, where teacher educators themselves have not been formally trained in language assessment, collaboration often constitutes the 'point of departure' for learning about language assessment.

The importance of collaboration in the development of language assessment competences has also been highlighted in previous LAL research. For example, Baker and Riches (2018), in a LAL project in Haiti, found that collaboration between teachers and assessment specialists facilitated teachers' LAL development during a series of workshops for the revision of English national examinations. Similarly, Harsch, Seyferth and Villa Larenas' (2021) LAL project with higher education language teachers in Germany identified collaboration between different stakeholder groups – teachers, coordinators, and researchers – as one of the most salient contributing factors to teachers' LAL development. Harsch et al (2021:333) observed that '[t]he processes of sharing experiences, exchanging expectations, forming common standards, developing materials and revising curricula collaboratively led to shared course descriptions, curricula and learning outcomes, and to end-of-course exams that reflect the shared values and aims'.

These findings also highlight the value of building a community which allows for cooperation and learning. Indeed, we found that communities were formed in the teacher educators' working contexts, in the shape of either formal or informal teams. In this regard, Xu and Brown (2016:158) have argued that '[p]articipation in community activities engages teachers in professional conversations about their assessment practices, offering opportunities to understand alternative thinking and practice of assessment, and allowing them to defend their own conceptions and negotiate their ideas with colleagues'. This, in turn, can change teachers' (or, in the case of

How do teacher educators learn about and teach language assessment?

this study, teacher educators’) daily practice of assessment and lead to an advancement of their assessment literacy.

The concept of *community of practice*, as embedded in the sociocultural theory of learning, refers to a social organisation where participation and learning occur. Wenger-Trayner and Wenger-Trayner (2015:1) define these communities as ‘groups of people who share a concern or a passion for something they do and learn how to do it better as they interact regularly’. They also explain that these communities come in a variety of forms: small, large, local, international, formally recognised, or informal, and even invisible. Our study found that communities of practice – both formal and informal – fostered collaboration among colleagues, thereby creating opportunities for learning about language assessment. According to Burns and Richards (2009:239), through dialogue and interaction, ‘teachers can come to better understand their own beliefs and knowledge as well as reshape these understandings through listening to the voices of others’. Similarly, Patton and Parker (2017:359) argue that ‘working together is effective ... the cooperative is a place where ideas belong to the group and where learning is promoted and valued’.

Additionally, our data revealed that communities of practice also offer opportunities for learning about language assessment through dialogic reflection. Mann and Walsh (2017:1) discuss reflective practice as ‘an important element in learning to be a competent teacher’. Our findings showed that reflective practice was greatly fostered when working towards the same goal within a social group (while individual reflective practice on language assessment did not occur systematically). In this regard, Mann and Walsh (2017:18) argue that ‘[w]hile independent and individual reflection does have its place, an over-emphasis on “lone reflection” devalues the importance of reflection as a series of collaborative processes’. Instead, Mann and Walsh (2017:189) stress the value of dialogic reflection as ‘a bottom-up, teacher-led, collaborative process entailing interaction, discussion and debate with another professional [which] can lead to professional learning’.

Learning from others has been reported in prior LAL research as a compensation strategy for insufficient language assessment training (Berry et al 2019, Vogt and Tsagari 2014). This seems a logical explanation for the Chilean EFLTE context, where teacher educators’ own language assessment training is virtually in-existent when taking on their role. However, this may not be the only reason for teacher educators to engage in communities of practice. Patton and Parker (2017:353) explain that communities of practice serve multiple purposes by providing ‘a professional learning process that is meaningful and relevant to individual members’. In line with this, we found that even more experienced teacher educators who were more confident in their language assessment practices saw great value in working

collaboratively; they felt that teamwork was beneficial for the exchange of ideas and further learning about language assessment. In other words, teacher educators' participation in community activities was not only a coping strategy for lack of language assessment training, but also perceived as a facilitating environment for continuous LAL development. Even in those universities where collaboration was not formally promoted, the teacher educators sought opportunities to engage with others in their practices. Indeed, Patton and Parker (2017:353) clarify that, through communities of practice, teachers 'seek to break down walls of solo practice ... and create spaces where faculty learn from and with each other, promoting professional growth'.

Cascading LAL: Teacher educators' LAL development of their student teachers

Our findings showed that, unless the EFLTE programme contained a dedicated language testing course, the LAL development which teacher educators offer to their student teachers is rather limited – both in terms of *what* the teacher educators teach and *how*.

Not surprisingly, the most LAL training was conducted by *language-testing-course* teacher educators (who teach a range of language assessment content, assess their student teachers in a variety of ways, continually and deliberately model their language assessment practices, and conduct reflection and practical analysis tasks with student teachers), but such courses are offered in only about half of Chilean EFLTE programmes. Alternatively, language assessment is supposed to be included in *teaching methodology* courses, but in practice, instruction on language assessment is limited in time and coverage on these courses, as they need to cover all aspects of language teaching theory and methodology.

In principle, the English acquisition courses in Chilean EFLTE programmes offer considerable potential for LAL training of student teachers. Tsagari and Vogt (2017:54) make the analogy 'test as [they were] tested', which means that student teachers are likely to repeat the testing practices they are exposed to, for example in their own EFL classes. Indeed, Graham (2005:610) explains that student teachers are 'profoundly influenced by their "apprenticeship of observation" ... they tend not to question assessment practices but instead implement plans that look like their own experience as students'. Thus, how *language* teacher educators (those conducting the English acquisition courses) assess their student teachers might greatly influence what the latter learn about language assessment and which practices they adopt later. In addition, given that language teacher educators spend the most instructional time with student teachers and thus student teachers are most exposed to these teacher educators' practices, the

How do teacher educators learn about and teach language assessment?

role of these teacher educators in the language assessment training of future EFL teachers can be capitalised upon.

In practice, however, we found that the *language* teacher educators' reports of their modelling and encouragement of reflective practice largely centred on language teaching, not assessment. One reason might be language teacher educators' own lack of language assessment training, which might result in low levels of language assessment self-efficacy and lack of self-identification as language assessors (see Villa Larenas and Brunfaut 2022). Consequently, they might not feel confident enough to purposely model language assessment practices for their student teachers or to have reflective discussions on their language assessment practices in their classes. Additionally, as teachers of English, language teacher educators might feel that it is not their direct responsibility to model language assessment practices and encourage awareness in student teachers, but that it is the responsibility of *testing* teacher educators.

Interestingly, *language* teacher educators did sometimes model language assessment practices, albeit unintended or unconsciously, when they engaged student teachers in their course assessment practices. Also, as language teacher educators constantly assess their student teachers' EFL proficiency in their courses, they effectively provide their student teachers with numerous examples of language tests, tasks and procedures. According to Singh and Richards (2009:204), '[a] challenge for anyone teaching [language teacher education] courses is how well the trainers' and the course's instructional practices model the kinds of learning opportunities and dispositions that teachers are encouraged to create in their own classrooms'. In the case of this study, the modelling of and reflection on language assessment practices seemed to be a challenge for the Chilean EFLTE context, and the few such strategies adopted in the English acquisition courses might widen the gap in LAL.

Implications and recommendations for LAL development in teacher education

While we know from prior research that there is a need for LAL development of pre- and in-service teachers, this study has shown an equal need for professional development of *teacher educators* in the area of language assessment. Although our study was set in the Chilean EFLTE context, based on our experiences and informal observations in a range of international contexts, we believe there are commonalities with other foreign language teacher education settings, and thus our recommendations might apply more widely.

Hadar and Brody (2016:58) remark that 'teacher educators' role in preparing the next generation of teachers is at the crux of educational

innovation and effective schooling'. By strengthening language assessment training in teacher education, new generations of language teachers with more extensive LAL will graduate each year. To achieve this, we recommend equipping teacher educators with the skills to purposefully model their language assessment practices and to stimulate reflective practices. Professional development efforts might thereby particularly concentrate on *language* teacher educators, since they constantly assess their student teachers' EFL proficiency and these assessments offer valuable opportunities for the LAL training of student teachers.

The LAL training these teacher educators then conduct with their student teachers could comprise not only sound language assessment practice models but also encourage student teachers' awareness of language assessments in their socioeducational contexts. In turn, as the student teachers graduate and move into the classroom, this might impact positively on teachers' local school realities. Furthermore, and importantly, it might benefit the teacher education context, as regular teachers typically end up pursuing teacher education careers. As Berry (2016:41) highlighted, it is commonly assumed that a good teacher will automatically be a good teacher educator, and '[c]onsequently, teacher educators tend to be "left alone" in terms of their professional preparation for their task'. However, as this study has shown for the Chilean context, language assessment expertise cannot be assumed of the current body of EFL teachers (who were not specifically trained in this area), and thus also not of those teachers 'drifting' (Berry 2016:40) into a teacher education role. Indeed, our study's findings suggest that student teachers' LAL training needs to go hand-in-hand with developing teacher educators' own LAL, not least to encourage confidence in their language assessment practices.

Thus, in conclusion, this study's findings suggest that the vital link to target in the language education chain, to enable a large LAL cascade in the language teaching community, is the teacher education context and both its stakeholders of teacher educators and student teachers.

References

- Baker, B and Riches, C (2018) The development of EFL examinations in Haiti: Collaboration and language assessment literacy development, *Language Testing* 35 (4), 557–581.
- Berry, A (2016) Teacher educators' professional learning: A necessary case of 'on your own?', in de Wever, B, Vanderlinde, R, Tuytens, M and Aelterman, A (Eds) *Professional Learning in Education. Challenges for Teacher Educators, Teachers and Student Teachers*, Ghent: Academia Press, 39–56.
- Berry, V, Sheehan, S and Munro, S (2019) What does language assessment literacy mean to teachers?, *ELT Journal* 73 (2), 113–123.
- Brindley, G (2001) Language assessment and professional development, in Elder, C, Brown, A, Hill, K, Iwashita, N, Lumley, T, McNamara, T and O'Loughlin, K (Eds) *Experimenting with Uncertainty: Essays in Honour of*

How do teacher educators learn about and teach language assessment?

- Alan Davies, *Studies in Language Testing Volume 11*, Cambridge: UCLES/Cambridge University Press, 126–136.
- Brown, J and Bailey, K (2008) Language testing courses: What are they in 2007?, *Language Testing* 25 (3), 349–383.
- Burns, A and Richards, J (2009) Second language teacher education through collaboration, in Burns, A and Richards, J (Eds) *Cambridge Guide to Second Language Teacher Education*, Cambridge: Cambridge University Press, 239–240.
- Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.
- Davies, A (2008) Textbook trends in teaching language testing, *Language Testing* 25 (3), 327–347.
- Freeman, D (2009) The scope of second language teacher education, in Burns, A and Richards, J (Eds) *Cambridge Guide to Second Language Teacher Education*, Cambridge: Cambridge University Press, 11–19.
- Gardner, S and Rea-Dickins, P (2001) Conglomeration or chameleon? Teachers' representations of language in the assessment of learners with English as an additional language, *Language Awareness* 10 (2–3), 161–177.
- Graham, P (2005) Classroom-based assessment: Changing knowledge and practice through preservice teacher education, *Teaching and Teacher Education* 21 (6), 607–621.
- Hadar, L and Brody, D (2016) Professional development for teacher educators in the communal context: Factors which promote and hinder learning, in de Wever, B, Vanderlinde, R, Tuytens, M and Aelterman, A (Eds) *Professional Learning in Education. Challenges for Teacher Educators, Teachers and Student Teachers*, Gent: Academia Press, 57–83.
- Harding, L and Kremmel, B (2016) Teacher assessment literacy and professional development, in Tsagari, D and Banerjee, J (Eds) *Handbook of Second Language Assessment*, Berlin: De Gruyter, 413–428.
- Harsch, C, Seyferth, S and Villa Larenas, S (2021) Evaluating a collaborative and responsive project to develop language assessment literacy, *Language Learning in Higher Education* 11 (2), 311–342.
- Johnson, K and Golombek, P (2011) *Research on Second Language Teacher Education: A Sociocultural Perspective on Professional Development*, London/New York: Routledge.
- Lam, R (2015) Language assessment training in Hong Kong: Implications for language assessment literacy, *Language Testing* 32 (2), 169–197.
- Levy-Vered, A and Alhija, F (2015) Modelling beginning teachers' assessment literacy: The contribution of training, self-efficacy, and conceptions of assessment, *Educational Research and Evaluation* 21 (5–6), 378–406.
- López, A and Bernal, R (2009) Language testing in Colombia: A call for more teacher education and teacher training in language assessment, *Profile* 11 (2), 55–70.
- Lortie, D (1975) *Schoolteacher: A Sociological Study*, Chicago: University of Chicago Press.
- Loughran, J (1995) Practising what I preach: Modelling reflective practice to student teachers, *Research in Science Education* 25 (4), 431–451.
- Loughran, J and Berry, A (2005) Modelling by teacher educators, *Teaching and Teacher Education* 21 (2), 193–203.
- Lunenberg, M, Korthagen, F and Swennen, A (2007) The teacher educator as a role model, *Teaching and Teacher Education* 23 (5), 586–601.

- Mackey, A and Gass, S (2016) *Second Language Research: Methodology and Design* (Second edition), New York: Taylor and Francis.
- Malderez, A (2009) Mentoring, in Burns, A and Richards, J (Eds) *Cambridge Guide to Second Language Teacher Education*, Cambridge: Cambridge University Press, 259–268.
- Mann, S and Walsh, S (2017) *Reflective Practice in English Language Teaching: Research-based Principles and Practices*, Abingdon: Taylor and Francis.
- Murray, J (2016) Beginning teacher educators: Working in higher education and schools, in Loughran, J and Hamilton, M L (Eds) *International Handbook of Teacher Education Volume 2*, New York: Springer, 35–70.
- Patton, K and Parker, M (2017) Teacher education communities of practice: More than a culture of collaboration, *Teaching and Teacher Education* 67, 351–360.
- Röehler, L and Cantlon, D (1997) Scaffolding: A powerful tool in social constructivist classrooms, in Hogan, K and Pressley, M (Eds) *Scaffolding Student Learning: Instructional Approaches and Issues*, Cambridge: Brookline Books, 6–42.
- Russell, T and Munby, H (1992) *Teachers and Teaching: From Classroom to Reflection*, London: Falmer Press.
- Singh, G and Richards, J (2009) Teaching and learning in the course room, in Burns, A and Richards, J (Eds) *Cambridge Guide to Second Language Teacher Education*, Cambridge: Cambridge University Press, 201–208.
- Tsagari, D and Vogt, K (2017) Assessment literacy of foreign language teachers around Europe: Research, challenges and future prospects, *Papers in Language Testing and Assessment* 6 (1), 41–63.
- Villa Larenas, S (2020) *An investigation of the language assessment literacy of teacher educators in Chile: Knowledge, practices, learning, beliefs, and context*, unpublished doctoral dissertation, Lancaster University.
- Villa Larenas, S and Brunfaut, T (2022) But who trains the language teacher educator who trains the language teacher? An empirical investigation of Chilean EFL teacher educators' language assessment literacy, *Language Testing* 40 (3), available online: doi.org/10.1177/02655322221134218
- Vogt, K and Tsagari, D (2014) Assessment literacy of foreign language teachers: Findings of a European study, *Language Assessment Quarterly* 11 (4), 374–402.
- Vygotsky, L (1978) *Mind in Society*, Cambridge: Harvard University Press.
- Wallace, M (1991) *Training Foreign Language Teachers: A Reflective Approach*, Cambridge: Cambridge University Press.
- Wenger-Trayner, E and Wenger-Trayner, B (2015) *Communities of Practice. A Brief Introduction*, available online: wenger-trayner.com/introduction-to-communities-of-practice/
- Woodward, T (1991) *Models and Metaphors in Language Teacher Training: Loop Input and Other Strategies*, Cambridge: Cambridge University Press.
- Xu, Y and Brown, G (2016) Teacher assessment literacy in practice: A reconceptualization, *Teaching and Teacher Education* 58, 149–162.

Appendix: Coding scheme

High-level codes	Subcodes
Teacher educators' backgrounds	<ul style="list-style-type: none"> • (Language) assessment training • general assessment • language assessment • no training • on the job • self-taught
Language assessment learning	<ul style="list-style-type: none"> • Teacher education <ul style="list-style-type: none"> • language teacher educator • methodology teacher educator • testing teacher educator • years in EFLTE • Community activities <ul style="list-style-type: none"> • confidence • cooperation • working alone • Reflective practice <ul style="list-style-type: none"> • about own assessment practices • about own teaching practices • about own needs • about materials shared during the interviews
Development of LAL with pre-service teachers	<ul style="list-style-type: none"> • Reflection <ul style="list-style-type: none"> • assessment practices • awareness • teaching practices • testing theory and practice connection • opinions on the language assessment training offered in EFLTE programmes • Modelling <ul style="list-style-type: none"> • assessment practices • language teaching • teacher skills • intended

5

Supporting K-12 educators' language assessment literacy via resources informed by validation frameworks

Ahyoung Alicia Kim

WIDA, University of Wisconsin-Madison, USA

Shireen Baghestani

Iowa State University, USA

David MacGregor

WIDA, University of Wisconsin-Madison, USA

Pauline Ho

WIDA, University of Wisconsin-Madison, USA

This chapter describes how an argument-based validation framework, specifically the Assessment Use Argument (AUA) by Bachman and Palmer (2010), can be used to inform the development of resources that are designed to support the language assessment literacy (LAL) of educators serving kindergarten to Grade 12 (K-12) English learners (ELs). It focuses on:

- The language assessment literacy needs of K-12 educators in relation to a federally mandated English language proficiency (ELP) assessment in the US
- The purpose of argument-based validation frameworks and their relevance for test stakeholders
- The development of AUA-informed resources for K-12 educators that serve ELs and their feedback on the resources
- A discussion of some useful principles for developing AUA-informed resources that are relevant and useful for K-12 educators

Este capítulo describe cómo se puede usar un marco de validación basado en el argumento, específicamente el argumento de uso de exámenes (Assessment Use Argument (AUA), Bachman y Palmer, 2010), para informar el desarrollo de recursos para apoyar la competencia en evaluación de idiomas (Language Assessment Literacy o LAL) de educadores que enseñan estudiantes de inglés (ELs) desde kínder hasta el grado 12 (K-12). Éste se enfoca en:

- Las necesidades de LAL de los educadores de los grados K-12 relativas a un examen del dominio del idioma inglés (ELP) exigido por el gobierno federal en los EE. UU
- El propósito de marcos de validación basados en el argumento y su importancia para las partes interesadas en los exámenes
- El desarrollo de recursos basados en el AUA para educadores de los grados K-12 que enseñan ELs, y sus comentarios sobre los recursos
- Una discusión de algunos principios útiles para el desarrollo de recursos basados en el AUA que son relevantes y útiles para educadores de los grados K-12

Introduction

Language assessment literacy (LAL) refers to the knowledge, skills and awareness stakeholders require to engage with language assessments appropriately (Harding and Kremmel 2016, Taylor 2013). For kindergarten to Grade 12 (K-12) educators in the US, the appropriate use of federally mandated English language proficiency (ELP) assessments is a key component of LAL. In recent years, there has been growing interest among language assessment researchers (Bachman and Damböck 2018, Bachman and Palmer 2010, Chapelle 2021, Schmidgall 2017, Schmidgall, Cid, Carter Grissom and Li 2021) in developing LAL resources for non-specialists using argument-based validation frameworks, an example being Bachman and Palmer's (2010) Assessment Use Argument (AUA). Frameworks such as the AUA provide a means of articulating the claims, warrants, and evidence supporting test score interpretation and use (Kane 2006, 2013), making them explicit not only for test developers but also various other test stakeholders, such as educators and policymakers. In theory, the information contained in such frameworks should enable stakeholders to not only evaluate the strength of a validity argument (Schmidgall et al 2021), but also gain a deeper understanding of how an assessment is intended to be used and why.

The need for LAL resources addressing test use is especially great for stakeholders in educational contexts, who are often responsible for

interpreting scores from assessments and using them for decision making, but may lack formal assessment training and/or have insufficient knowledge about the assessments they use (Baker 2016, Baker, Tsushima and Wang 2014, O’Loughlin 2011, 2015). This is particularly true in US K-12 settings, where educators are required to administer an annual ELP assessment to ELs as part of the federal law (Every Student Succeeds Act of 2015) to ensure that ELs are provided with the necessary language instruction to promote their English language development (Bailey and Carroll 2015, Sireci and Faulkner-Bond 2015). ELP assessment developers have a responsibility to support this effort by providing educators (e.g., administrators, EL teachers, and general education teachers) with meaningful data and resources about students’ English language proficiency that can inform placement of students, resource allocation, and programming.

In this chapter, we describe our process of developing AUA-informed resources to support the LAL of K-12 EL educators in their use of ACCESS for ELLs (hereafter ACCESS). ACCESS is an ELP assessment developed by WIDA¹ in collaboration with the Center for Applied Linguistics. ACCESS is used across 41 US states, territories, and federal agencies for making various decisions such as placement of students into language instruction educational programs (e.g., EL/ESL programs), monitoring students’ year-to-year progress in their English language proficiency, and program evaluation. The test consists of four domains (listening, speaking, reading, and writing) and is anchored in the five WIDA English Language Development (ELD) Standards: social and instructional language plus the language of the four content areas of language arts, math, science, and social studies (WIDA 2012). It is important that educators be provided with resources that support their LAL to understand the purpose of ACCESS, as well as the meaning and use of its scores. Resources informed by an argument-based validation framework, such as an AUA (Bachman and Palmer 2010), may be especially useful for this purpose by providing examples of how the test is intended to be used for best student outcomes, along with explicit statements about unintended uses of the test.

Language assessment literacy

Following Taylor’s (2013) observation that, regarding LAL, ‘[n]ot everyone needs to know or be able to do everything to the same level’ (2013:409), LAL research often seeks to identify LAL profiles for different stakeholder

1 WIDA at the University of Wisconsin-Madison is an organization that supports K-12 multilingual learners in the US by creating English Language Development Standards and English language proficiency assessments, and offering professional learning opportunities to educators.

groups in terms of what competences stakeholders already possess or may need to develop given the ways they engage with language assessments. Such research has tended to focus on pre- or in-service language teachers and their needs and abilities related to classroom assessment (Berry, Sheehan and Munro 2019, Crusan, Plakans and Gebril 2016, Fulcher 2012, Malone 2013, Vogt, Tzagari and Spanoudis 2020, Yan, Zhang and Fan 2018). In recent years more research has been conducted on users of high-stakes, standardized assessments, namely college admissions personnel, including those responsible for setting language test score requirements and making admissions decisions (e.g., Baker 2016, Baker et al 2014, Deygers and Malone 2019, O'Loughlin 2011, 2015).

Despite the integral role ELP assessments play in the K-12 education of ELs, there has been little research on LAL profiles of K-12 educators with regard to standardized assessments. A recent exception to this is a study by Kim, Chapman, Kondo and Wilmes (2020) which investigated K-12 EL educators' use of ACCESS score reports. Participants were asked to share which methods of score reporting used in the ACCESS Individual Student Report were most helpful for understanding their students' test performance. The study sheds light on K-12 EL educators' LAL profile in terms of both their preferences and familiarity with different score reporting methods. Rather than suggesting that educators have deficiencies in LAL, however, the findings provide insight into how EL educators think about proficiency scores and suggest ways testing organizations can adapt score reports to provide the greatest benefit for their stakeholders. The current study seeks to add to our understanding of K-12 EL educators' LAL profile by exploring the ways they engage with and may benefit from test resources informed by an argument-based validation framework.

Argument-based validation frameworks

Argument-based validation (Kane 1992, 2006, 2013) involves creating a validity argument – explicit documentation of the logic behind the case for using an assessment for a particular purpose. In theory, such documentation should enable test stakeholders to evaluate the logic and assumptions regarding the assessment put forward by the test developers and to ensure that the assessment is used in an appropriate way. Different validation frameworks, or ways of structuring validity arguments, have been proposed in the field, such as Kane's (2013) Interpretation/Use Argument (IUA) framework or Bachman and Palmer's (2010) Assessment Use Argument (AUA). An AUA consists of four claims regarding the *assessment records*, score *interpretations*, *decisions* made based on the interpretations, and the *consequences* of these decisions. Bachman and Palmer (2010) provide a number of suggested qualities for each of these four claims: consistency

(of assessment records); meaningful, impartial, generalizable, relevant, and sufficient (interpretations); equity and values-sensitivity (of decisions); and beneficence (of consequences). The claims are linked through an inferential reasoning process originally formulated by Toulmin (1958, Kane 2006), often depicted through the use of arrows to represent inferences and boxes to represent claims.

Two notable efforts to employ the AUA framework in efforts to support stakeholders' LAL include the work by Bachman and Damböck (2018) and Schmidgall et al (2021). In their book, *Language Assessment for Classroom Teachers*, Bachman and Damböck (2018) explain the test development process for classroom teachers using an AUA. When used in this way, the AUA becomes an articulation of the positive qualities the assessment will have, and therefore guides the test development process. Similarly, Schmidgall et al (2021) provide an AUA for the TOEIC Bridge test designed to enable test consumers to 'better evaluate whether the TOEIC Bridge tests are appropriate for their situation' (2021:1). In their paper, Schmidgall et al (2021) articulate the four claims in the AUA, along with their justification (warrants and backing), in technical yet simplified detail. Their presentation of the AUA reflects the flexibility of the AUA framework; although the AUA itself remains the same, it can be presented more or less technically (i.e. in more or less detail) depending on the audience for whom it is intended (Bachman and Palmer 2010). In a similar way, WIDA has adapted the AUA to develop the validation framework for ACCESS, which guides ACCESS test development and rationale behind the test use (Kim, MacGregor and Cook 2018). The AUA for ACCESS is a lengthy technical document with claims and various supporting evidence, which may not be suitable for K-12 educators, but its core concepts could support the LAL of K-12 educators. For this reason, we have chosen the AUA as the basis to inform the LAL resources described in this chapter.

Research questions

In this chapter², we discuss a project aimed at creating accessible and relevant AUA-informed test-related resources for K-12 educators to support their LAL to appropriately use test scores and promote beneficial outcomes for ELs. The resources consist of a short document and a video. This chapter

2 This research was part of a larger study that involved the development of an AUA for ACCESS and also AUA-informed educator resources to support their LAL in using ACCESS. The former is technical in nature with full-fledged information on various claims and evidence to support them, and is geared towards an audience with advanced LAL. On the other hand, the latter is short and concise for educators who may not necessarily have training in language testing. In this chapter, we present the findings on developing AUA-informed prototypes of educator resources.

describes our journey in developing the prototypes of these two resources. Specifically, we explore what types of content we found to be suitable (and not suitable) for our audience of educators; the findings were subsequently used to revise the prototypes into final versions (WIDA 2022a, 2022b) and these are described in later sections. The main focus of this chapter is the process of discovery we experienced along the way, with each stage leading us to new questions and ideas. For instance, although our original goal was to develop LAL resources for K-12 EL educators, we discovered that general educators, including content or mainstream teachers, could also benefit from the types of resources we planned to develop. As a result, our first research question, which originally addressed only EL educators, was revised to address all K-12 educators. The study was guided by the following two research questions (RQs):

1. What are K-12 educators' LAL needs for understanding and using ELP assessments?
 - a. What implications could this have for creating LAL resources for ELP assessments?
2. How could an argument-based validation framework inform the development of LAL resources for K-12 educators using ELP assessments?

Methodology

The methodology for this study consisted of two phases. In Phase 1, a prototype of a short AUA-informed document (hereafter 'assessment use document') was developed, presented to educators, and discussed as part of a focus group. In Phase 2, a short video (hereafter 'assessment use video') was created based on the assessment use document and educator feedback from Phase 1. The video was then presented to educators and discussed as part of a focus group.

Phase 1: Assessment use document

The assessment use document created for Phase 1 was three pages long, presenting a simplified, more educator-friendly version of the full AUA for ACCESS (Kim et al 2018). The full version of the AUA includes claims and detailed supporting evidence (e.g., research, documentation, policy). By contrast, the educator-friendly version we created borrowed the core concepts of the AUA (i.e., the four main claims of *assessment records*, *interpretations*, *decisions* and *consequences*) without going into much detail on the supporting evidence. As seen in Appendix A, the assessment use document contained a table listing the intended decisions and consequences

of using ACCESS adapted from the existing AUA, a diagram representing the Toulmin-style logic of the AUA for ACCESS, and a table providing the actual claims and types of support alluded to in the diagram.

The first focus group was conducted online for 70 minutes and recorded. It was facilitated by two of the authors and attended by eight EL educators from different US states that use ACCESS. The participants had an average of 20 years of experience as educators and 14 years working specifically with ELs. Half of the educators were working as EL teachers at the school level, while the other half worked as EL program coordinators or administrators at the district level.

The educators received the assessment use document by email and were asked to familiarize themselves with it prior to the meeting. The meeting began with the facilitators briefly reviewing the document (about 10 minutes), followed by an hour-long question and answer session. The questions³ were organized around four themes: (1) educators' use of ACCESS and associated challenges and concerns regarding the test; (2) potential use of the assessment use document; (3) suggestions for improving its content; and (4) format of the assessment use document (see Appendix B). The findings from Phase 1 are reported in the Results section.

Phase 2: Assessment use video

The original plan was to employ the assessment use document as foundational content for developing a short video. In fact, the educator feedback concerning the assessment use document proposed a more engaging resource, so for Phase 2 we created the assessment use video which became a prototype for a later, more fully developed version of the video that is now available on the WIDA website (WIDA 2022b). The prototype assessment use video was a little over four minutes and provided practical guidance on how educators can use ACCESS to create a positive educational impact (e.g., placement and programming decisions). Using non-technical terms, it narrated the importance of well-designed tests and test administration, so that scores could be used for making accurate interpretations and appropriate decisions, which would lead to beneficial outcomes for ELs. Like the earlier assessment use document, the video is structured around the four basic claims in an AUA – assessment records (test administration), interpretations, decisions, and consequences. Although the video does not provide a formal argument,

³ When conducting focus groups, the research team did not find it necessary to have a separate session to examine educators' LAL needs. This was partly because WIDA regularly engages with educators when developing assessments and resources. However, the first focus group question generated helpful information for refining the AUA-informed resources.

we believe it retains enough of a connection to the basic AUA structure to call it 'AUA-informed.'

The second focus group was also conducted online and recorded. Because it was conducted as part of a larger meeting with multiple agenda topics, the discussion of the video only lasted about 15 minutes. The meeting was attended by two of the authors and seven educators from states that use ACCESS. The educators served ELs as district EL directors, coaches, or coordinators, and they had a mean of 8.6 years of experience with WIDA assessments.

The educators received a link to the video via email prior to the meeting and were asked to view it in advance. During the focus group, the educators were presented with three questions concerning: (1) potential use of the assessment use video; (2) suggestions for improving the content of the video; and (3) resources to supplement the video, if needed (see Appendix B).

Data analysis

Each focus group was video-recorded, transcribed, and analyzed using an open-coding procedure (Saldaña 2016). Two members of the research team independently read through the transcript data and generated a preliminary list of words and phrases relevant to the questions asked. For this we used participants' own words, also known as *in vivo* coding (see Appendix C for sample coding). These words and phrases were then used to identify themes in the participants' responses and create an initial coding scheme, which was then refined through discussions among all members of the research team. Once the coding scheme was finalized, two researchers coded the focus group transcript independently. They then met to discuss any code disagreements until they reached full agreement.

Results

In relation to RQ1, we first present findings regarding our perceptions of educators' LAL needs related to the use of ACCESS. Because participants in Phase 2 were not asked about their LAL needs directly, these findings mostly reflect data from Phase 1. Regarding RQ2, we describe what we learned about how AUA-informed resources could be used to support the LAL of K-12 EL educators, particularly in terms of their use of ACCESS.

Findings on RQ1: LAL needs related to the use of ACCESS scores

K-12 EL educators in the Phase 1 focus group identified several ways they used ACCESS. These findings are presented briefly as they informed

the AUA-informed resources we developed. Determining placement and language services for ELs was the most frequently mentioned use of ACCESS scores; another common use was to inform instruction. In conjunction with ACCESS scores, WIDA Can Do Descriptors (2016), which describe what learners at various score levels are able to do with language, were often mentioned as a useful resource when planning instruction. A third common use of ACCESS scores was to facilitate communication between EL teachers and other stakeholders, such as parents and content teachers, about the language needs of ELs. Participants mentioned that it was helpful to refer to the scores while talking to students and/or parents about a child's learning goals. Educators mentioned additional ways of using ACCESS scores: to monitor patterns and trends in students' performance, to determine where to allocate resources, and to meet federal accountability requirements. Not surprisingly, most of these latter comments came from district-level EL administrators or coordinators rather than EL teachers at the school level.

In terms of their concerns when using ACCESS scores to make decisions, and the additional information that would be helpful when using the scores, educators' responses could be grouped into three themes: (1) lack of familiarity with ACCESS among content and special education teachers; (2) score report information and its interpretation; and (3) concerns related to EL administration and policy. The first theme refers to the overall lack of familiarity with ACCESS among content and special education teachers, resulting in teachers under-using ACCESS scores to improve EL instruction. This theme also came up repeatedly in the Phase 2 focus group. For instance, Dianne (pseudonym) commented that there was a 'lack of familiarity [among non-EL educators] with the assessment and its results' and a lack of 'understanding of how they can take that information and integrate that into the work that they're doing with students on a regular basis'. It is not surprising that EL educators expressed concerns regarding special education teachers, considering approximately 10% of ELs are identified with mild cognitive disabilities and are required to take ACCESS (meanwhile, students with severe cognitive disabilities take Alternate ACCESS). In the second focus group, Melissa (pseudonym) mentioned the 'pushback' against requiring students with disabilities to take ACCESS:

We get a lot of pushback from students with disabilities. How will they ever pass the test? ... that is really misunderstood in our special education world, that they see it as, well, if they have reading disability, they'll never be able to gain language proficiency, which I know is wrong. But maybe that's a separate video.

In addition, Nicole, a participant in the second focus group, said 'Anything we can do to help our colleagues in [Special Education (SpEd)] understand

how ACCESS data can help them better serve ELs who are also [students with disabilities (SWD)] would be useful'. Such comments suggest that more information needs to be provided by districts, states, or WIDA to help special education teachers use ACCESS scores effectively.

The second theme included concerns about score reports and their interpretation. Educators shared comments regarding the type of information included in score reports. They indicated wanting more granular information on students' specific strengths and weaknesses rather than just an overall score. One educator also indicated the need for more information on interpreting students' growth. In addition, educators raised concerns about interpreting test scores, particularly for the speaking domain. According to Michelle, 'When we switched over from ACCESS 1.0 [paper test] to ACCESS 2.0 [online platform] our scores definitely shifted, with the speaking scores just totally tanking under ACCESS 2.0. There is some skepticism about why those speaking scores are so low'. Some speculated that less motivated students may bring down score averages.

The third theme included concerns related to EL administration and policy. For example, some states use ACCESS for teacher evaluation, even though the test is not intended to be used this way. This policy therefore places an unfair burden on educators, as will be discussed later on. In addition, there were concerns regarding the lack of consistency in EL classification policy across states, and the challenges of upholding the federal requirement to administer ACCESS to SWDs. Finally, one person mentioned the difficulty both EL and general education teachers have distinguishing between the English language arts assessment (a content assessment administered to students in mainstream classes and ELs) and ACCESS (a language proficiency assessment only required for ELs). This last point was also raised in the second focus group.

Summary of findings on RQ1

Overall, EL educators indicated using ACCESS in a variety of ways, which is consistent with the literature (Sireci and Faulkner-Bond 2015). EL educators generally felt confident with their ability to use ACCESS scores (e.g., placing students, communicating with and about ELs' language development). Nevertheless, they highlighted several areas of concern as presented above, suggesting that score use may not always be straightforward. Furthermore, their comments revealed that many of their concerns were not with their own score use or that of fellow EL educators, but with that of general educators (e.g., mainstream/content teachers, administrators) and also special education teachers. This suggests that LAL resources about ACCESS score use may be equally if not more important for general educators as for EL educators. The fact that general educators lack knowledge about how ACCESS scores can inform instruction echoes concerns about the lack of

training for content teachers on effective instruction for ELs (e.g., Bailey and Heritage 2017).

Another finding that became clear was that developing a single resource to address all the LAL needs of educators was unfeasible given the wide range of needs from various audiences. In addition to the EL educators who we spoke to, general educators may require resources that would familiarize them with ACCESS and guide them on ways to use it to support ELs. In addition, numerous topics were raised including concerns about low speaking test scores, the difference between ACCESS and English language arts assessments, and best practices for testing SWDs. Thus, we proceed to the next section with the caveat that although the needs of educators are many, the AUA-informed resources we developed only addressed a limited number of those needs.

Findings on RQ2: The nature and format of ‘educator-friendly’ AUA-informed resources

Focus group findings from Phases 1 and 2 indicate the content and format educators prefer in the AUA-informed resources. Specifically, educators reported preferences for (1) non-technical and accessible information, (2) practical information that addresses educators’ needs, and (3) visually appealing resources that multiple audiences could use to gain familiarity with ACCESS.

Educators wanted resources that contained non-technical and easily accessible information. This is reflected in their feedback on the assessment use document and the assessment use video; while educators generally liked the video, they had difficulty comprehending the document. The assessment use document (Appendix A) contained a summary of the intended uses and consequences for ACCESS and explained how an AUA is structured through Toulmin-style logic, consisting of claims and supports. Due to the technical nature of the document, involving jargon, educators perceived the document to be inappropriate as an educator resource. In addition, the text was dense with a lot of information. Thus, educators remarked that it was ‘text heavy’, ‘not user friendly’, and difficult to understand. Michelle for instance commented that ‘I had to read it several times, to get my head wrapped around what I think you folks are trying to get me to understand. So I’m still not quite totally firm on it.’ Participants also suggested making the document shorter and using bullet points rather than long paragraphs. Additionally, they suggested creating an interactive resource, PowerPoint, or video.

A couple of educators also felt the assessment use document might be more appropriate for those in higher-level administrative positions. For instance, Linda commented that:

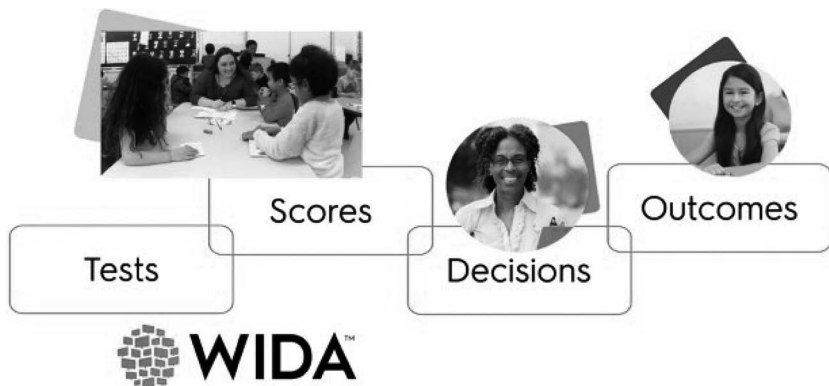
[The document] wouldn't be something I would share with teachers or even parents. It seems like something that the state could use, like the Department of Education. Maybe administrators and schools or even district superintendent too, especially those who may not have an understanding of ACCESS, need to see the importance, but it does seem not as user friendly on more of that day-to-day basis.

In contrast to the assessment use document, the assessment use video was quite well received. Instead of a Toulmin-style chart, the video was organized around the four core claims of the AUA, which were simplified into one-word headings: Tests, Scores, Decisions, and Outcomes. The importance of each 'stage' in the testing process, including the role educators play in upholding quality at each stage, was narrated verbally with supporting images (see Figure 1). When asked to comment on what they liked about the video, three educators mentioned that it was clear, concise, and visually appealing.

Although the assessment use document was generally viewed as being overly technical and opaque, participants remarked that one aspect they did like was the table listing the intended uses and consequences of ACCESS (see Appendix A). This echoed Michelle's comment that, 'teachers, and to a certain extent the principals at the school level, they want more practical [information]. They want to know what it is that they can do *now*.'

In response to this feedback, the video was made to address educators directly by discussing how they played a role in ensuring success at each stage of the testing process. Feedback on the video indicated that educators appreciated how it was very explicit in saying how ACCESS was intended to be used, but they discussed also wanting information on how scores should *not* be used. The latter type of information is rarely discussed in existing ACCESS resources developed by WIDA. This is partly because state and

Figure 1 Image from assessment use video



local education agencies determine how they will use test scores in addition to following federal guidelines. However, local policies that are not aligned with the intended test use could potentially lead to negative consequences. Melissa commented that examples of what ‘ACCESS *is* intended to be used for’ needed to be balanced with the examples of ‘what it is *not* intended to be used for’. In keeping with educators’ needs, Melissa commented that:

there is such confusion between how [ACCESS] is different from language arts content standards, and [I’m] wondering if there is a way to embed a bit more ... I think there’s a place for really being explicit about how it’s different from those content standards.

Most of the educators in the second focus group said they could see educators using the video as an introduction to ACCESS to become well informed about the test. Only one educator (Leslie) did not feel the video would be useful in her context, and this was because she taught in a dual immersion program where non-EL teachers were already familiar with ACCESS: ‘We have a lot of ELs, and most of our teachers I think have a pretty good understanding as do administrators, already. So, while I see a really great use for this video, I think some of our staff might need to go deeper.’

When asked what they liked about the video, three participants mentioned the fact that it could be used with a range of audiences which included EL teachers, content teachers, special education teachers, principals, and other administrators (see Table 1). Educators repeatedly mentioned how important it was to raise awareness among general educators about how and why ACCESS is used, and they seemed to believe that the video could achieve that. All in all, the video seemed to strike a chord with focus group participants who believed it could reach a critical audience in need of more information about ACCESS.

Summary of findings on RQ2

In summary, educators preferred the assessment use video over the assessment use document due to it being non-technical and accessible; including practical information that addresses educator needs; and being visually appealing for a wide range of audiences. The fact that educators could see themselves using the video speaks to its ability to address educators’ needs. Despite their enthusiasm about the video, participants shared lingering issues that they hoped WIDA would address, including clarifying the difference between ACCESS and language arts content standards, and helping special education teachers make better use of ACCESS to serve students with disabilities. While some of these issues may be outside the scope of the current assessment use document or video, it may be possible in the future to create more targeted

Table 1 Potential audiences for the assessment use video as suggested in focus groups

Audience	Purpose the video would serve	Representative quotations
Principals and administrators	Guide appropriate use of ACCESS scores	'One of the things [I] particularly like ... was when you were talking about the scores, and what they should not be used for ... I think it would be particularly important for our administrators to see that part of it.' (Sandra)
EL teachers	Not specified	'I might share it with ELL specialists and then have them share that with the teachers and administrators in their buildings ... as part of a short introduction to ACCESS testing.' (Christopher)
Content teachers	Introduce the purpose of ACCESS testing	'[It might] be effective to use with classroom teachers to help them become aware of [ACCESS]. Because it seems like we pull students out of their classroom to do this test once a year. And sometimes they're not aware of what the test is for ... and I think the video does a good job of encouraging [stakeholders] to ... empower themselves to know how the information is used.' (Christopher)
Special education teachers	Guide the appropriate interpretation and use of ACCESS scores	'This would be a great intro for our SpEd staff. They really just equate ACCESS to Academic Standardized tests, and accommodations on the Individualized Education Programs (IEPs) are all over the place. At meetings, they don't understand the scores and find it hard to read the data and use it appropriately.' (Allison)

AUA-informed resources, which address these and other concerns raised by stakeholders. Future videos could also provide more information on test use tailored towards varying educator groups with different knowledge and experience with ELs.

Discussion and conclusion

The study described in this chapter arose from a perceived need to communicate to educators the appropriate and inappropriate uses of ACCESS. Our initial plan was to couch that communication in terms of the AUA already being developed for the test. However, in our discussions with educators, it became clear that, while the AUA could provide an overall

framework in creating materials for educators, the terminology and detail in the AUA make it difficult for educators to access. Thus, we pivoted from developing *AUA-based* materials to *AUA-informed* materials, using the lessons we learned from our first focus group to create materials more tailored to educators' needs.

Our discussions with educators in this study showed awareness of the different claims of the AUA – particularly the claims on interpretations, decisions, and consequences – even if that awareness was expressed in different terms. Regarding interpretations, educators focused on the need for general education and special education teachers to better understand the meaning of ACCESS test results, for example, the difficulty that many have in differentiating between the construct of reading proficiency as measured by ACCESS and reading comprehension as measured by content tests. In relation to decisions, they mentioned the use of ACCESS scores in placement of ELs, and informing curriculum and instruction (in conjunction with the WIDA Can Do Descriptors). Finally, regarding consequences, they expressed concerns about misuse of ACCESS scores, especially the use of test scores for teacher evaluation.

The participating educators found the prototypes of AUA-informed resources (i.e. the document and the video) potentially useful for a wide variety of audiences, including EL educators, content teachers, special education teachers, school and district-level administrators, and state/federal administrators. Future resources should be developed to support the specific LAL needs of these different educator groups. More research will need to be conducted to better understand the needs of these groups, but based on the results presented here some possibilities include: (1) helping EL educators interpret ACCESS scores; helping content teachers understand the need to adjust instruction to the needs of ELs (Bailey and Heritage 2017); (2) helping language arts teachers understand the difference between the test constructs for language arts assessments and ACCESS; (3) giving special education teachers information about the test content and helping them understand how to adjust their instruction to ELs and recommend appropriate accommodations (Abedi 2021); and (4) helping EL educators in administrative roles understand intended and unintended uses of ACCESS scores.

As a result of our conversations with educators in Phases 1 and 2, we subsequently revised the assessment use document prototype into a one-page flyer, designed to provide educators with information about the intended and unintended uses of ACCESS (see Appendix D; WIDA 2022a). The flyer is organized around the steps of the AUA (scores, interpretations, decisions, and uses), without necessarily using the vocabulary of the AUA or describing the particulars of the arguments. Instead, it focuses on the different roles that local educators, state administrators, and WIDA play in those steps. The intent of the flyer is to provide that information in a short, easily

understandable, useful format. In addition, the assessment video prototype underwent revisions to enhance its clarity of the content (WIDA 2022b). While the one-page assessment use document could be easily disseminated via email and also printed, the video resource can be more appealing to stakeholders who prefer such a format.

In the context of K-12 education, all teachers are, at least to some degree, language teachers. From the perspective of educators in the focus groups, reaching content-area and special education teachers is equally as urgent as reaching language teachers. Additionally, educators underscored the importance of communicating to administrators the intended and, perhaps more importantly, unintended uses of an assessment. For all of these purposes, we hope both the updated version of the assessment use document and video could serve as effective and appealing vehicles.

In developing resources based on the AUA, we learned some important lessons on how to communicate important information from complex, technical documents. First, avoid jargon. While educators may be familiar with the concepts encompassed by a validation argument, they cannot be expected to have had extensive training in assessment. Second, keep it short. Educators are often pressed for time, and so any document longer than one or two pages, or a video longer than a few minutes, will run the risk of being ignored. Third, make it attractive. In creating both the video and the handout, we worked with colleagues with expertise in communication and graphic design to create materials that communicated our message in easily digested, short, and attractive ways. And finally, make it relevant. Educators we spoke with were enthusiastic about the assessment use video because they saw an actual use and a real need for it. Our findings strongly resonate with literature on creating assessment resources (e.g., score reports) – that consideration of an audience's needs, knowledge, and attitudes is essential in designing test-related documents (Zapata-Rivera and Katz 2014).

It is important to mention that our data comes exclusively from what a limited group of EL educators discussed in the focus groups. Collecting more data from a larger sample of EL educators may therefore provide a broader range of perspectives. In addition, one should be cautious of relying too much on EL educators' perception of how other educators use ACCESS. Future research could benefit from conducting additional focus groups, consisting of content and special education teachers. This would allow us to better understand the needs of varying groups of educators. In this chapter we have shown how resources informed by an AUA and adapted for a particular audience can be used to convey important and useful information about test score interpretation and use, and to support educators' LAL. While our first attempt at creating a resource (i.e., assessment use document prototype) that would be engaging and useful for educators proved somewhat unsuccessful, we were able to glean invaluable feedback in our conversations

with educators, and to use that feedback to craft a more accessible assessment use video along with a more focused and accessible assessment use document (WIDA 2022a, 2022b). This experience underscores the importance of engaging with the target audience of stakeholders when creating materials intended to communicate important information about the interpretation, uses, and consequences of a testing program, and illustrates a methodology for translating technical concepts into language that is relatable to a range of key stakeholders who may not necessarily be testing specialists.

References

- Abedi, J (2021) Accommodations and universal design, in Fulcher, G and Harding, L (Eds) *The Routledge Handbook of Language Testing*, New York: Routledge, 306–321.
- Bachman, L F and Palmer, A (2010) *Language Assessment in Practice*, Oxford: Oxford University Press.
- Bachman, L F and Damböck, B (2018) *Language Assessment for Classroom Teachers*, Oxford: Oxford University Press.
- Bailey, A L and Carroll, P E (2015) Assessment of English language learners in the era of new academic content standards, *Review of Research in Education* 39 (1), 253–294.
- Bailey, A L and Heritage, M (2017) Imperatives for teacher education: Findings from studies of effective teaching for English language learners, in Peters, M A, Cowie, B and Menter, I (Eds) *A Companion to Research in Teacher Education*, New York: Springer, 697–712.
- Baker, B A (2016) Language assessment literacy as professional competence: The case of Canadian admissions decision makers, *Canadian Journal of Applied Linguistics* 19 (1), 63–83.
- Baker, B A, Tsushima, R and Wang, S (2014) Investigating language assessment literacy: Collaboration between assessment specialists and Canadian university admissions officers, *Language Learning in Higher Education* 4 (1), 137–157.
- Berry, V, Sheehan, S and Munro, S (2019) What does language assessment literacy mean to teachers?, *ELT Journal* 73 (2), 113–123.
- Chapelle, C A (2021) *Argument-based Validation in Testing and Assessment*, Thousand Oaks: SAGE Publications Inc.
- Crusan, D, Plakans, L and Gebril, A (2016) Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices, *Assessing Writing* 28, 43–56.
- Deygers, B and Malone, M E (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.
- Harding, L and Kremmel, B (2016) Teacher assessment literacy and professional development, in Tsagari, D and Banerjee, J (Eds) *The Handbook of Second Language Assessment*, Berlin: De Gruyter Mouton, 413–427.
- Kane, M T (1992) An argument-based approach to validity, *Psychological Bulletin* 112 (3), 527–535.

- Kane, M T (2006) Validation, in Brennan, R B (Ed) *Educational Measurement* (Fourth edition), Westport: American Council on Education and Praeger, 17–64.
- Kane, M T (2013) Validating the interpretations and uses of test scores, *Journal of Educational Measurement* 50 (1), 1–73.
- Kim, A A, Chapman, M, Kondo, A and Wilmes, C (2020) Examining the assessment literacy required for interpreting score reports: A focus on educators of K–12 English learners, *Language Testing* 37 (1), 54–75.
- Kim, A A, MacGregor, D and Cook, H G (2018) *The Assessment Use Argument for ACCESS for ELLs 2.0 (WIDA Internal Report)*, Wisconsin: University of Wisconsin–Madison.
- Malone, M E (2013) The essentials of assessment literacy: Contrasts between testers and users, *Language Testing* 30 (3), 329–344.
- O’Loughlin, K (2011) The interpretation and use of proficiency test scores in university selection: How valid and ethical are they?, *Language Assessment Quarterly* 8 (2), 146–160.
- O’Loughlin, K (2015) ‘But isn’t IELTS the most trustworthy?’: English language assessment for entry into higher education, in Ata, A and Kostogriz, A (Eds) *International Education and Cultural-linguistic Experiences of International Students in Australia*, Samford Valley: Australian Academic Press, 181–194.
- Saldaña, J (2016) *The Coding Manual for Qualitative Researchers* (Third edition), London: SAGE.
- Schmidgall, J (2017) *Articulating and Evaluating Validity Arguments for the TOEIC® Tests*, RR-17-51, Princeton: Educational Testing Service.
- Schmidgall, J, Cid, J, Carter Grissom, E and Li, L (2021) Making the case for the quality and use of a new language proficiency assessment: Validity argument for the redesigned TOEIC Bridge® tests, *ETS Research Report Series 2021* (1), 1–22.
- Sireci, S G and Faulkner-Bond, M (2015) Promoting validity in the assessment of English learners, *Review of Research in Education* 39 (1), 215–252.
- Taylor, L (2013) Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections, *Language Testing* 30 (3), 403–412.
- Toulmin, S E (1958) *The Uses of Argument*, Cambridge: Cambridge University Press.
- Vogt, K, Tsagari, D and Spanoudis, G (2020) What do teachers think they want? A comparative study of in-service language teachers’ beliefs on LAL training needs, *Language Assessment Quarterly* 17 (4), 386–409.
- WIDA (2012) *2012 Amplification of the English Language Development Standards: Kindergarten–Grade 12*, WIDA: University of Wisconsin–Madison.
- WIDA (2016) *WIDA Can Do Descriptors: Key Uses Edition, Grades 9–12*, WIDA: University of Wisconsin–Madison.
- WIDA (2022a) *Using ACCESS for ELLs To Promote Beneficial Outcomes for English Learners*, WIDA: University of Wisconsin–Madison.
- WIDA (2022b) *Using ACCESS for ELLs To Promote Beneficial Outcomes for English Learners* [Video], available online: wida.wisc.edu/resources/using-access-ells-promote-beneficial-outcomes-english-learners-0
- Yan, X, Zhang, C and Fan, J J (2018) ‘Assessment knowledge is important, but ...’: How contextual and experiential factors mediate assessment practice and training needs of language teachers, *System* 74, 158–168.
- Zapata-Rivera, J D and Katz, I R (2014) Keeping your audience in mind: Applying audience analysis to the design of interactive score reports, *Assessment in Education: Principles, Policy & Practice* 21 (4), 442–463.

Appendix A: Assessment use document (prototype)

Assessment Use Argument for ACCESS for ELLs and WIDA Screener

The purpose of ACCESS for ELLs (hereafter ACCESS) and WIDA Screener is to provide an accurate assessment of the academic English language proficiency of English learners (ELs). Scores on ACCESS and WIDA Screener are intended to assist educators in making decisions about ELs’ English support needs and to ultimately improve the quality of education for ELs. Table 1 provides a list of the decisions ACCESS and WIDA Screener are intended to be used for and the consequences that should arise from appropriate use of these scores.

Table 1 Intended uses (decisions) and consequences of using ACCESS and WIDA Screener

Intended decisions	Intended consequences
<p><i>Decisions about K-12 ELs</i></p> <ul style="list-style-type: none"> • Classification as EL or non-EL • Placement into appropriate LIEP • Reclassification <p><i>Stakeholders involved in some or all decisions: K-12 school- or district-level administrators and educators, families of ELs</i></p>	<ul style="list-style-type: none"> • K-12 ELs receive the appropriate language instruction educational program (LIEP) instruction and support they need to enhance their academic English language proficiency • ELs will be able to meaningfully participate in content classes where the language of instruction is only English when they are reclassified as ELs
<p><i>Decisions about programming</i></p> <ul style="list-style-type: none"> • Allocation of resources for LIEPs • Identification of schools needing additional support • Decisions to change instruction <p><i>Stakeholders involved in some or all decisions: K-12 school- or district-level administrators, schools (in small districts), local education agencies, teachers</i></p>	<ul style="list-style-type: none"> • School-level EL administrators and educators can use ELs’ information regarding their English language proficiency to enhance their LIEPs • EL administrators at local education agencies can determine the effectiveness of LIEPs and seek ways to further improve them

An argument can be made that the use of ACCESS and WIDA Screener leads to improved quality of education for ELs. However, this positive outcome (consequence) is dependent upon the appropriate use of test scores and the validity and reliability of these scores. Figure 1 provides a diagram of an Assessment Use Argument (AUA) – an argument justifying the use of test scores for a particular purpose (see Table 2). In the diagram, the rectangles represent the major claims that are used to build the AUA (note that due to

space, the full claims are provided in Table 2). The claims are numbered from 1 to 4, beginning with Intended Consequences. This numbering is due to the fact that consequences are the first consideration in the test development process. However, test users and educators should read the diagram from bottom to top, starting with Claim 4 and ending with Claim 1.

The upward-pointing arrows above the other rectangles indicate that the claim below the arrow supports the claim above it. The standalone arrows represent additional support for the claims above them. For example, in order for Claim 3 ('test scores are an indicator of academic English language ability') to be valid, it must be supported by Claim 4. However, there would also need to be additional support in order for Claim 3 to be valid. In other words, Claim 4 is necessary but not sufficient support for Claim 3.

Figure 1 Diagram of an Assessment Use Argument

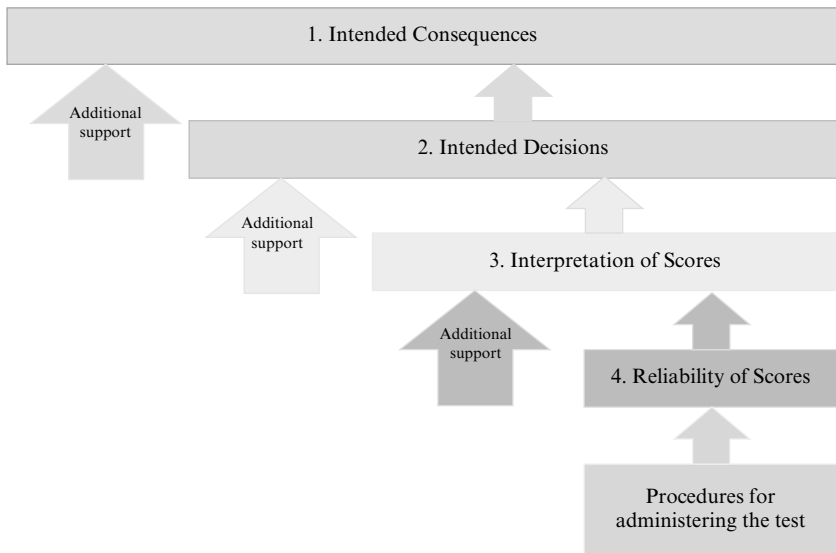


Table 2 Major claims of the AUA and their support (should be read starting with Claim 4)

Claim		Support
1. Intended Consequences:	The consequences of the decisions made using ACCESS and WIDA Screener scores are beneficial to stakeholders.	<ul style="list-style-type: none"> • Claim 2: Decisions based on test scores are fair and align with the interests of stakeholders. Additional support: <ul style="list-style-type: none"> • Procedures are followed to prevent or mitigate detrimental consequences. For example, test records are kept confidential to prevent the release of private information. • Stakeholders can use and benefit from the information score reports provide because they are easy to understand and delivered in a timely manner. • Claim 3: Test scores indicate the extent to which students have the academic English language proficiency to succeed in mainstream U.S. public school classrooms. Therefore, decisions are based on accurate and appropriate information.
2. Intended Decisions:	Decisions based on test scores are fair and align with the interests of stakeholders.	<p>Additional support:</p> <ul style="list-style-type: none"> • The minimum scores (cut scores) used to classify ELs generally result in accurate classification. • The federal Every Student Succeeds Act (ESSA) law, and each state's and district's guidelines are adhered to when making EL classification and programming decisions. • Claim 4: Test performance is collected and scored using fair and consistent procedures.
3. Interpretation of Scores:	Test scores indicate, with a high degree of accuracy, the extent to which students have the academic English language proficiency to succeed in mainstream U.S. public school classrooms.	<p>Additional support:</p> <ul style="list-style-type: none"> • The material tested on ACCESS and WIDA Screener, as well as the scoring criteria, are representative of both the content and range of academic language used in U.S. public schools. • Attributes of the test or test delivery which could disadvantage test-takers are minimized. This helps to ensure that students are able to demonstrate their actual/true academic English language proficiency.
4. Reliability of Scores:	Test performance is collected and scored using fair and consistent procedures.	<ul style="list-style-type: none"> • ACCESS administrative procedures are followed consistently across different testing venues, from year to year, and for all student groups. • Human raters must be certified and regularly monitored to ensure that their ratings are unbiased, consistent over time, and consistent with those of other raters. Regardless of who rates a response or when it is rated, the score should be the same. • Multiple-choice items are scored by computer. Quality control procedures are in place to ensure that human error does not compromise the accuracy of computer scoring. • Different test forms are equivalent in terms of length, delivery, difficulty, and in the range of language skills which are assessed. Therefore, all students are tested in the same way regardless of which version of the test they take.

Appendix B: Focus group questions

Phase 1 questions

1. Use of ACCESS scores
 - a. How do you use scores from ACCESS? What's the most important use of test results?
 - b. When using ACCESS scores to make decisions, what concerns do you have?
 - c. What kind of information do you need to better understand the use of ACCESS and its scores? (e.g., making decisions using ACCESS scores)
2. Potential use of assessment use document for ACCESS
 - a. What do you see yourself using the document for? How would you use it?
 - b. How could this document change/improve the way you use ACCESS?
3. Suggestions for improving the content of the assessment use document
 - a. Would you like to see any additional content added to the document?
 - b. What information in the document is unclear or unnecessary?
 - c. Is there terminology that is difficult to understand?
4. Suggestions for improving the format of the assessment use document
 - a. How could we improve the format of the document?
 - b. Do you prefer to have the document in another format (e.g., video tutorial, webpage, etc.)
 - c. Do you have any suggestions on the length of the document? What is the ideal length?

Phase 2 questions

1. Do you see yourself using the assessment use video? If so, for what purpose?
2. Is everything in the video accessible to the audience you have in mind? Is anything unclear?
3. If WIDA created a handout to go with this video, would you use it? If so, would a one-page summary (mostly text), an infographic (mostly images, minimal text), or something in between best communicate the information to the audience you have in mind?

Appendix C: Examples of coded segments

RI: Okay, okay sorry I was confused with my mute and unmute button for a moment. So how do we use the scores from ACCESS? The most important use, to me, is class placement to make sure that we get a correct fast placement or level. Oh, I don't know how they do it in elementary. Being a middle school, we are their language arts teachers. So, and the second, that's the secondary model. So, making sure that we use those scores, and we look at those literacy scores and all the pieces of it to make sure that they're correctly placed in the class for the next school years so that they get, so, we start with where they're at. So, we don't give them too easy classes, we don't want to put them in over their head. So, it's really important to be able to look at those scores and use them to place kids in classes. Some of the other uses, of course, are to let kids know how they're doing, and how they're gaining language. I use them to look at my own instruction. Other teachers that I work with, and other EL teachers that I work with do the same thing. And kind of look and see if there's any consistency. Maybe you're consistently high somewhere. You're like, okay, we really did great with the reading this year. Or maybe you look at it and say, okay, the speaking was lower. We--; maybe we didn't do enough speaking practice. So, I kind of use it as a gauge for myself as well, to look for patterns of where I might be doing well or need to improve any own instruction. We can look at we look at it as a district also to see how we're doing that way. But to me, the most important thing is just making sure we're really precise with that class placement for the student the next year. So that student has the best opportunity to grow the most.

When I make decisions the concerns I've had, and I've shared these with you, and I've talked to my EL manager. We've really had some issues with the recording of the speaking. And so, I really sometimes I look at the overall score, the speaking score seemed particularly low. I look at that overall score and I wonder how legitimate that is. So that's the questions that I run into. And that's a technology issue on our part that we need to solve. But that is a concern. And what the thing I would like to know, I, I'd like to be able to drill the data deeper than I can right now. So, for example, if they are, let's say a student does have a low speaking score. Why? Are they not saying enough? Are they not giving completed sentences? Are they ... where's the hang-up? What do they need to improve it? Are they not using enough content vocabulary, or are they not using the correct grammar structures? Is there any way to find out just really exactly where they're struggling on the test because sometimes those scores surprise me? I, a student I know read for me very, very well in class and then just totally tanks the reading, where is that mismatch? Where, you know, so, that would be some information that I would love to have is, how can I drill deeper with that data?

<p>Author Use of ACCESS Scores – Placements & Services</p> <p>...</p> <p>Reply</p>
<p>Author Use of ACCESS Scores – Communicate about students' ED</p> <p>...</p> <p>Reply</p>
<p>Author Use of ACCESS Scores – Inform curricula</p> <p>...</p> <p>Reply</p>
<p>Author Use of ACCESS Scores – Monitoring patterns and trends</p> <p>...</p> <p>Reply</p>
<p>Author Use of ACCESS Scores – Placements & Services</p> <p>...</p> <p>Reply</p>
<p>Author Concerns with using ACCESS Scores – Score accuracy – validity</p> <p>...</p> <p>Reply</p>
<p>Author Additional info wanted – Granular student ED info</p> <p>...</p> <p>Reply</p>

Appendix D: Assessment use document (final version)

Using ACCESS for ELLs to Promote Beneficial Outcomes for English Learners

How to use ACCESS for ELLs scores to achieve positive outcomes for English learners

	Test scores	Interpretations	Decisions
School/ district-level educators	Ensure schools have the necessary training and technology to administer ACCESS. Follow state and WIDA administration requirements and guidelines.	Understand what ACCESS scores mean and what a student can do with language. Communicate the meaning of ACCESS scores to content teachers, families, and students.	Consistently apply decision criteria to all students. Inform instruction by helping educators make decisions about language programs and resource allocation.
State-level administrators	Ensure districts and schools have the necessary resources and guidance to administer ACCESS.	Understand what ACCESS scores mean and how they might be used to inform school and district improvement.	Create state-level policy on placement, progress monitoring, program evaluation, and reclassification.
WIDA	Rely on educator input to develop a test with fair and relevant content.	Provide clear guidelines for interpreting ACCESS scores.	Provide clear guidelines and useful resources for making decisions using ACCESS scores.

Positive and Intended Outcomes

- ✓ Everyone understands students' year-to-year English language proficiency development.
- ✓ Students receive language support at the level they need.
- ✓ Students can meaningfully participate and are supported in content classes conducted in English.
- ✓ Students who are proficient in English participate in content classes without need for further English language support.
- ✓ School-level educators receive targeted and sufficient resources to provide high-quality instruction to English learners.

Negative and Unintended Outcomes

- ✗ Test results are used as measure of ability in English language arts.
- ✗ Students are placed into special education or gifted education programs based on test performance.
- ✗ Teacher effectiveness is judged based on student test performance.



WIDATM

WIDA is housed within the Wisconsin Center for Education Research at the University of Wisconsin-Madison.
© 2022 The Board of Regents of the University of Wisconsin System, on behalf of WIDA

6

Enhancing teachers' and test-takers' assessment literacy? Insights from test preparation for the digital TestDaF

Sonja Zimmermann

Leska Schwarz

Anja Peters

Günther Depner

g.a.s.t.¹, TestDaF-Institut, Germany

This chapter aims at evaluating a new approach for test preparation that is based on the principles of learning-oriented assessment (LOA). The chapter focuses on:

- How the principles of LOA informed the development of test preparation materials for a high-stakes standardised proficiency test for admission purposes in Germany
- The opportunities for teachers and learners to develop language assessment literacy (LAL) by using these materials
- Language teachers' perceptions of this approach for test preparation

Der vorliegende Beitrag hat zum Ziel, einen neuen Ansatz zur Prüfungsvorbereitung zu evaluieren, der auf den Prinzipien des *Learning-oriented assessment* (LOA) aufbaut. Im Fokus des Kapitels stehen:

- Die Berücksichtigung dieser Prinzipien bei der Entwicklung von Vorbereitungsmaterialien für einen standardisierten High-Stakes-Test für die Hochschulzulassung in Deutschland
- Die Möglichkeiten für Lehrende und Lernende, durch die Verwendung dieser Materialien ihre *Language Assessment Literacy*

¹ g.a.s.t. is the acronym for *Gesellschaft für akademische Studienvorbereitung und Testentwicklung e. V.* (Society for Academic Study Preparation and Test Development).

(LAL), also die Expertise über theoretische und praktische Grundlagen des Prüfens und Testens, zu entwickeln

- Die Meinungen von Lehrkräften zu einem solchen Ansatz für die Prüfungsvorbereitung

Introduction

When preparing for a large-scale language test, learners and teachers need to have some kind of language assessment literacy (LAL) which can be broadly defined as the knowledge, skills and principles related to language assessment (Davies 2008). While these three components have remained constant in the ongoing theoretical and research-based discussions about LAL, other aspects continue to be an issue of debate, in particular how LAL can be best developed among different stakeholders (Inbar-Lourie 2013, Taylor 2013). In the context of language learning in general and test preparation in particular, one of the central stakeholder groups is classroom teachers, and a great amount of the existing research has focused on the kind of knowledge, skills and principles LAL for this specific group should entail (e.g. Fulcher 2012, Malone 2013, Popham 2009, Vogt and Tsagari 2014).

One area of language teaching where LAL takes a central role – not only for teachers but also for learners – is test preparation classes, as these aim at building knowledge and awareness of assessment in prospective test-takers, an important stakeholder group whose needs have not been the focus of LAL studies yet (Butler, Peng and Lee 2021). Even though not explicitly mentioning LAL, studies on test preparation practices and preferences might provide some insights into the knowledge of test-takers about language tests, their attitude towards the assessment and their views on useful test preparation. For example, Knoch and her colleagues (Knoch, Huisman, Elder, Kong and McKenna 2020) investigated the effect of self-access test preparation activities on test performance of repeat test-takers in the context of the Pearson Test of English (Academic). Using Messick's (1982) classification of types of test preparation, Knoch et al found that score improvement for the speaking section was mainly related to construct-irrelevant test preparation methods. The interviews with the test-takers also revealed that many of them engaged in activities like test familiarisation, and only a few engaged in language learning, i.e. in ways to improve their general language skills. O'Sullivan, Dunn and Berry (2021) looked into the preferences of learners for test preparation activities in different educational, political and social contexts, also taking into account test-taker characteristics like gender and age. Using questionnaire data from learners in countries of the Middle East, East and Southeast Asia as well as Central and South America, the results of this study indicate that

test preparation approaches and preferences are similar across the globe. Respondents regarded downloadable practice material and exercises on their mobile devices as most helpful, while they rated face-to-face classes and blended learning settings as less helpful, clearly indicating the preference for self-accessed test preparation, mainly focusing on test-taking skills.

While these preferences with a focus on test familiarisation and test-taking skills may be understandable from a learner's perspective, activities such as working through model tests rarely allow for language learning in the sense of improving general language skills. In comparison, approaches that follow principles of learning-oriented assessment (LOA) facilitate a greater integration of test preparation and language learning in test preparation classes. According to Carless (2007:59f), the following three principles apply for LOA: 1) assessment tasks should be conceptualised as learning tasks and as such mirror real-life activities 'to stimulate sound learning practices'; 2) students should be involved in the assessment, 'actively in engaging with criteria, quality, their own and/or peers' performance'; and 3) feedback 'should be timely and forward-looking'. By fostering a close link between test preparation activities and language learning settings, LOA can be a useful approach to ensure that 'test preparation provides an authentic learning experience' (Jones and Saville 2016:79), and hence has a positive washback on test preparation classes. Such an approach is also promoted by Gebriel (2018). Based on Messick (1996), Gebriel recommends that test preparation activities should be integrated into language learning settings and reflect real-life situations of target language use (TLU). This approach to test preparation will promote 'real learning', that is, the development of construct-relevant language skills, as opposed to teaching-to-the-test. He lays out how such a learning-oriented approach to test preparation can be achieved, stressing the importance of teachers' LAL, i.e. their understanding of the test construct, the task requirements and how this can be linked to beneficial classroom activities.

The current study therefore sets out to investigate the link between a learning-oriented approach to test preparation and the opportunities for learners and teachers to develop LAL in the context of such specific language learning settings. To do so, the study takes a closer look at preparation materials for the digital Test of German as a Foreign Language (*Test Deutsch als Fremdsprache – TestDaF*). Based on interviews with teachers of test preparation classes, the materials will be evaluated with respect to their usefulness for test preparation purposes as well as their potential for LAL development.

Context of the study

TestDaF is a standardised proficiency test that is officially recognised for language admission to Higher Education (HE) in Germany (Hochschulrektorenkonferenz and Kultusministerkonferenz 2020). Thus, it

caters for an ever increasing group of international applicants to German universities: in 2021, 2.9 million students attended 420 German universities, 14.1% of whom were international students (Hochschulrektorenkonferenz 2021). Since its introduction in 2001, the paper-based version of the TestDaF has established itself as the 'go-to-assessment' (Norris and Drackert 2018) for admission purposes. With more than 446,000 candidates so far, it is 'the German language test for university admission with the most participants' (g.a.s.t. 2020:2). In late 2020, a new digital version of TestDaF became operational. The test development process applied international standards for quality assurance (Kecker and Eckes 2022, Kecker, Zimmermann and Eckes 2022), and was based on a comprehensive needs analysis with data from more than 120 university lecturers and over 1,300 international students (Arras 2012, Marks 2015).





The digital TestDaF consists of four sections: reading, listening, writing, and speaking (see Figure 1). It contains a total of 23 task types, including integrated tasks (Cumming 2013, Plakans 2013). These test tasks cover eight competencies that were identified as essential for first year students in any degree programme at a German HE institution:

- recognise and reflect the positions/opinions of others and contrast them (if necessary) with one's own
- distinguish between personal opinion and factual argument
- take a stance (in speaking and writing)
- recognise and express differences or similarities
- recognise and express causal relationships
- comprehend graphics and reproduce information from them in one's own words
- take notes and, where required, use these for the production of written or spoken texts
- process information from summaries and produce summaries.

Test-takers' performances in each of the four components are related to one of three TestDaF levels (TestDaF-Niveaus – TDN) of language proficiency – TDN 3, 4 or 5 – which correspond to the Common European Framework of Reference for Languages (CEFR) levels B2 to C1. The successful completion at a specified minimum level (TDN 4)² serves as evidence of the German language skills required to gain admission to almost any discipline and degree course at universities and HE institutions in Germany.

2 See the TestDaF website for more information on admission regulations and the relation of TestDaF levels and the CEFR: www.testdaf.de/de/hochschulen/der-testdaf-und-hochschulen/nachweis-der-deutschenkenntnisse-fuer-das-studium/

Figure 1 Format of the digital TestDaF

	Lesen <i>Reading</i>	7 Aufgaben 34 Items 7 tasks 34 items	ca. 55 Min. <i>approx. 55 min.</i>
	Hören <i>Listening</i>	7 Aufgaben 30 Items 7 tasks 30 items	ca. 40 Min. <i>approx. 40 min.</i>
	Schreiben <i>Writing</i>	2 Aufgaben 2 tasks	ca. 60 Min. <i>approx. 60 min.</i>
	Sprechen <i>Speaking</i>	7 Aufgaben 7 tasks	ca. 35 Min. <i>approx. 35 min.</i>

The digital test environment allows the inclusion of media, such as video files, thereby illustrating the communicative situation for test-takers and increasing the authenticity of the test tasks. As in the paper-based TestDaF, topics covered in the test include the natural sciences, humanities and engineering. However, the tasks require no expert knowledge of individual fields of study and their terminologies.

Before launching the digital TestDaF in 2020, several piloting phases with international study applicants in preparatory language classes had been conducted. The piloting of the new test tasks was accompanied by test-taker questionnaires which included closed items as well as items giving participants the opportunity to add comments as free text. The questions focused on time allotment, usability, the clarity of task instructions and response layout, and included questions such as the following: Was the time for task completion sufficient? How well did you understand the task instructions? How user-friendly did you find the graphical interface? Moreover, participants were asked about aspects relating to the processing of the different task types and challenges encountered, e.g. if they found the use of pictures or videos helpful, or if they had trouble using the keyboard.

As Zimmermann (2021) showed, participants regarded the digital TestDaF as very user-friendly in terms of usability and the tools provided,

such as text enlargement and zoom functions. Naturally, the questionnaire data also revealed some challenges encountered by the participants, such as the large variety of test tasks or the single play of audios and videos.

Overall, test-takers reported that what they experienced as most challenging were the communicative competencies that are being elicited by the test tasks. For example, some test-takers commented that they had never written a summary before or that they found it difficult to paraphrase information and take notes while listening to an audio. This feedback was of some concern as these are exactly the competencies that language learners need in order to successfully participate in academic studies in HE institutions in Germany. It was therefore decided that any test preparation material or activities for the digital TestDaF would need to address more than just simple test-taking skills. Hence, competency-based teaching materials with a learning-oriented approach were developed for use in preparation classes at TestDaF test centres. These materials were also intended to specifically inform teachers about the test construct and the underlying approach for test preparation, not only to promote 'appropriate forms of test preparation' (Green 2017:115), but also to enhance their LAL.

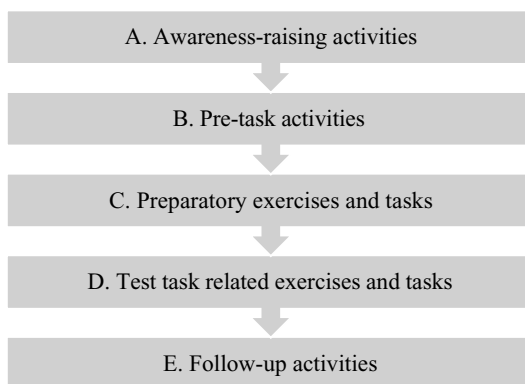
The next section includes a more comprehensive description of the TestDaF test preparation materials that were developed and how these are related to the principles of LOA.

Preparation materials

The preparation materials for the digital TestDaF were developed in cooperation with experienced teachers of German as a foreign language who, in the past, had taught test preparation courses for the paper-based TestDaF. The aim of involving teachers in the development of the materials was to benefit from their comprehensive expertise in language teaching in the context of preparatory language classes in Germany. The materials are based on a theoretical concept (g.a.s.t. 2019) which builds on two central aspects: a) focusing on the competencies mentioned above that underlie the test tasks across the different test sections, rather than focusing on single task requirements; and b) raising test-takers' awareness of the requirements of the test tasks and how these are related to the TLU domain. The latter especially offers potential for teachers' and learners' LAL development.

The course materials consist of six units, covering various topics from different domains and, in total, around 100 hours of instruction. They aim at reinforcing the acquisition and further development of communicative competencies that are required in academic contexts. Each course unit is divided into 5–10 sub-units consisting of five consecutive steps (see Figure 2).

Awareness-raising activities (Step A) enable learners to become aware of their individual resources. This step includes opportunities for reflection that

Figure 2 Structure of the learning units

reveal specific needs from the TLU situation. For instance, learners are asked to reflect on previous listening or reading experiences in German, or to evaluate the effectiveness of specific writing strategies they already use. Therefore, learners raise their awareness of academically relevant communicative situations and the skills and strategies required by those situations. Pre-task activities (Step B) include, for example, analysing images, discussing quotations, or developing a hypothesis about a given topic. These activities are followed by steps for approaching the actual exercises and tasks presented in Step C. Learners approach the target task through short exercises, e.g. capturing the main ideas of a text or completing information from a listening text. Step D refers to the development and training of specific competencies that are related to the actual test tasks of the digital TestDaF, e.g. activating receptive and productive vocabulary, implementing a fast and efficient word identification, weighing and arranging information when reading and listening or using appropriate linguistic registers when writing an argumentative text. For this, learners activate and initiate those cognitive processes that are part of the respective competencies. Follow-up activities (Step E) denote the last steps of a learning unit. Learners individually reflect on what they have learned. They evaluate the strategies they have used (retrospectively) or discuss further task-solving procedures (prospectively). The evaluation can, for example, be conducted in the form of a self-created checklist.

The five steps are supplemented by learning tips, providing students with information on learning strategies and additional ideas for independent further learning activities outside the classroom setting.

Taking into account the above stated principles of LOA (assessment tasks as learning tasks, student involvement, and feedback as 'feedforward'), the previously described preparation materials entail relevant aspects of this approach:

- Since the test tasks of the digital TestDaF are already closely linked to the requirements of the TLU domain, classroom activities that are related to these tasks automatically mirror real-life activities in the HE study context.
- The awareness-raising activities in Step A and some of the follow-up activities in Step E expect learners to engage with the requirements of communicative situations in academia, i.e. with the criteria they have to meet to successfully participate in these situations, or to reflect on their performance and/or successful strategies, hence developing self-evaluation skills.
- The material promotes the discussion among peers within the classroom, and peer feedback can be seen as timely and forward-looking feedback students engage with.

Teachers can use the provided materials as a toolbox to plan and carry out their preparation courses tailor-made for specific groups of learners by choosing, expanding and/or modifying the materials, and thus facilitating the integration into local language learning settings as suggested by Gebril (2018).

In order to encourage teachers to use the learning-oriented approach for preparing students for the digital TestDaF, the material is accompanied by a teacher's guide. This additional document supports teachers by explaining the desired learning outcomes of each individual task, suggesting opportunities for independent study and, where appropriate, pointing to similarities and differences between the learning tasks and test tasks. In this way, the guide fosters transparency about the construct underlying the test tasks of the digital TestDaF and enables teachers to make students aware of the link between the test tasks and the TLU requirements.

The design of the material as a toolbox and the accompanying teachers' guide primarily have the potential to develop teachers' LAL. In addition, engaging learners in activities that require them to reflect on *why* and *how* they are being assessed, the materials offer opportunities for the LAL development of learners as well.

In order to see how this learning-oriented approach was perceived by language teachers and to what extent in their opinion it can enhance test-takers' LAL, an exploratory study was conducted to evaluate the test preparation material for the digital TestDaF.

Research questions

Initially, the overall research aim was to evaluate the test preparation material from two different perspectives: that of the learner and of the teacher. More specifically, semi-structured interviews with both groups were

expected to shed light on the perception of the underlying theoretical concept of raising test-takers' awareness for a more learning-oriented approach to test preparation.

Due to the COVID-19 pandemic, many live examinations and preparation courses for the digital TestDaF had to be cancelled in 2020–2022, which reduced the number of possible interview participants drastically. Hence, only data from a small group of language teachers in Germany and Russia who used the material in their preparation courses were collected. The preparation courses differed with regard to mode (online vs. face-to-face), duration (four weeks vs. one semester) and the number of participants (between 5 and 20).

The final study centred around the following two research questions (RQs):

RQ1: To what extent do the teachers possess LAL with respect to the digital TestDaF? To be more specific: To what degree does their knowledge of the requirements in the TLU domain, i.e. university, enable them to link test preparation to study preparation?

RQ2: How do the teachers perceive the learning-oriented approach of the preparation material for the digital TestDaF? In what ways do they think this could contribute to enhancing test-takers' LAL?

Participants

Interviews were conducted with six teachers of preparation courses for the digital TestDaF, three of whom were also involved in the development of the preparation materials. All of the teachers were female and had extensive teaching experience (ranging from 6 to 20 years) in preparing international study applicants for taking up their studies in Germany, with a special focus on preparing them for the paper-based TestDaF. Most of them were teaching at language centres at German universities. At the time of the interview, some were teaching a test preparation class for the digital TestDaF for the first time, using the material provided, others had already finished a course in which they had used the material.

Data collection and analysis

To answer the above-mentioned RQs, semi-structured interviews in German were conducted via Microsoft Teams by two researchers. They lasted between 35 and 90 minutes and were recorded in order to facilitate transcription. The interview guide covered the following areas:

- experience in preparing learners for the paper-based or digital TestDaF, study preparation and the use and development of learning materials (RQ1)
- awareness of the challenges for international study applicants in preparing for taking up their studies in HE institutions in Germany (RQ1)
- knowledge about the construct underlying the test tasks of the digital TestDaF (RQ1)
- use of and opinions on the test preparation material (RQ2)
- challenges for those who were involved in the development of the materials (RQ1 and 2).

The recordings of the interviews were transcribed verbatim by student research assistants, applying conventions for simple transcripts (Dresing and Pehl 2018). The transcription was checked for quality by the researchers and the transcripts were then analysed in NVivo 12, following a structural coding approach (Saldaña 2016). For this chapter, parts of the interviews were translated into English.

Findings

This section reports findings in relation to the overall research aim of this study: the evaluation of the learning-oriented approach to test preparation and of its potential to contribute to teachers' and test-takers' LAL development. With regard to the first RQ, one aspect that emerged from the interviews and gave evidence of teachers' LAL was remarks relating to learners' attitudes towards test preparation. Teachers reported that learners often regarded test preparation as separate from more general study preparation activities. They were more interested in teaching-to-the-test activities, i.e. activities that focus on task types represented in the test (see also Zimmermann 2009). Such observations led some teachers to reflect on differences between the paper-based and the digital TestDaF. For example, one teacher pointed out that in preparation for the paper-based version of the test learners could be easily 'drilled':

Den papierbasierten [TestDaF] fand ich jetzt nicht so schwierig. [...] Und bei bestimmten Aufgaben konnte man sie sehr gut drillen, was ja nicht eigentlich Sinn und Zweck ist und das fällt jetzt beim digitalen TestDaF weg. Da sehe ich erstmal eine Umstellung, klar, auch bei der Vorbereitung, dass man eben nicht mehr drillen kann – was aber auch gut ist./*The paper-based [TestDaF], I didn't find that hard. [...] Certain tasks just allowed for drilling exercises which was actually not the intent and the purpose – something that comes to an end with the digital TestDaF. That is a change, obviously, no more drilling exercises in test preparation – which is good.*

Recognising that ‘drilling’ learners does not prepare them for the requirements of language use in academia, teachers were hoping that the introduction of the digital TestDaF would establish a closer link between test preparation activities and TLU, especially with the integrated task types:

[Bezüglich der] neuen Aufgaben kann ich sagen, dass mir die integrierten Aufgaben auch gut gefallen, weil das eine Möglichkeit bietet in der Vorbereitung also aus Lehrkraft-Perspektive, aber auch aus Teilnehmenden-Perspektive, das hat wirklich dann eine Authentizität, was im Studium so erwartet wird./*And regarding the new tasks I can say that I really like the integrated tasks since this is an opportunity in preparing for the test, not only from a teacher’s perspective, but also from a learner perspective; this is authentic, this is what is required at university.*

Teachers expressed that from their perspective, not only the integrated tasks, but also the test tasks in general mirrored the TLU requirements better than the tasks of the paper-based version of the TestDaF, as another teacher said in regard to the listening section:

Und ja, ich war überrascht, aber ich war positiv überrascht, da ich diese Aufgaben ja kenne aus dem Studium, die Anforderungen kenne. Und insofern war ich eigentlich auch positiv beeindruckt, dass es [...] eher wirklich einen sehr lebendigen, einen sehr quasi-authentischen Bezug hatte, da steht jemand und zeigt eine Folie und weist auch darauf hin, da stehen zwei Leute und sprechen miteinander und teilen durchaus unterschiedliche Ansichten, und das auch mimisch und gestisch./*And yes, I was surprised, but positively surprised since I knew these tasks, the requirements from university. And therefore I was positively impressed that there now is [...] a vibrant, an authentic relation; there is [a video of] someone showing a presentation slide and referring to it; there are two people talking, having different views, also by facial expressions and gesture.*

Comments like these about the authenticity of task types were found across all interviews, which shows that the teachers had a good understanding of the importance of task authenticity for test preparation practices. Furthermore, they recognised that the aim should be to make test-takers aware of this link between test preparation and TLU, and thereby expand learners’ LAL:

Ich würde mich freuen, wenn die Teilnehmenden das auch so sehen wie wir. Also es ist der Tenor generell, der neue TestDaF [ist] näher am Studium. Wenn die das auch so sehen würden, fände ich das super./*I would be happy if the test-takers saw it like we do. The general view is that the new TestDaF is more closely linked to communicative activities at university. If they also see it like this, I would be happy.*

Overall, teachers considered the missing link between test preparation and study preparation one of the main challenges in preparing international students for German university. With the introduction of the new test format of the digital TestDaF and a more authentic representation of the TLU requirements in the test tasks, they were hoping for a positive washback. Such reflections give evidence of the teachers' competency in matters relating to assessment, such as test task design and how the task design (e.g. task authenticity) impacts test preparation.

The second RQ looked into the teachers' perceptions of the learning-oriented approach of the preparation material for the digital TestDaF, and to what extent they thought it might contribute to enhancing test-takers' LAL. Overall, the material and the approach taken were well received by the teachers; in their opinion, the introduction and the teacher's guide were especially helpful. Even though all teachers had many years of experience in teaching, they acknowledged that a learning-oriented approach was something new for them in the context of test preparation:

Selbst für erfahrene Lehrkräfte ist die Vorbereitung auf den digitalen TestDaF ja schon was Neues und das ist eine große Hilfe, ne? Man hat das zwar auch im Lehrwerk, aber dann wird ja nicht so direkt herangeführt. Also auch die Tipps, die gegeben wurden, also Lehr- und Lerntipps, ... klar, das hat man im Hinterkopf, aber das fand ich gut, dass es auch nochmal aufgeführt wurde. Dass man es auch einfach nicht vergisst, gerade so beim ersten oder zweiten Mal, wenn man sich da so entlang hangelt./*Yes, but this is something new, even for experienced teachers test preparation for the digital TestDaF is new, and therefore this helps. You can find this approach in course books, but not that detailed or systematic. The specific teaching and learning tips you get, of course you have them in mind, but it helps to have them listed there, so that you don't forget about them, especially when you make your way along for the first or second time.*

The interviews revealed that teachers used the material to different extents. While some of them worked with single units, following all steps from A to E as outlined above, others made use of only some activities that were thematically linked to other materials they worked with. One teacher reported that she chose the material depending on the difficulties learners reported with specific test tasks and the underlying competencies or cognitive processes. The individual use of the test preparation material in relation to external conditions like course duration and the needs of learners showed that teachers made use of the material as intended. The adaptation to local language learning settings is also evidence of teachers' LAL development.

Sometimes the material was perceived as too detailed, so that teachers left out some of the steps (A–E) in the learning units. Interestingly enough, only

one teacher reported that she mainly skipped the awareness-raising activities (Step A) due to time constraints, falling back on teaching test-taking skills instead:

Ich habe es eher weggelassen. Vor allem weil ich auch ein bisschen unter Zeitdruck stehe, also viereinhalb Unterrichtseinheiten in der Woche für den ganzen digitalen TestDaF ist nicht viel und deswegen, natürlich führe ich ein, aber ich halte mich da jetzt nicht unbedingt an diese Wahrnehmungslenkung, was hier angegeben ist oder vorgeschlagen ist in den Unterlagen, sondern ich mache das vielleicht kürzer oder auch anders, aber konzentriere mich dann doch eher auf die Aufgaben./I left that out. Mainly because of time pressure, 4.5 hours per week for preparing for the digital TestDaF is not much, and therefore, of course I introduce a topic, but I don't follow the material when it comes to the awareness raising activities. I shorten that or do it differently, but in the end, I focus more on the tasks.

However, this was an exception and all other teachers stressed that these awareness-raising activities were very important for them to include in their classes. Some of them even made extra room for these activities and the exchange between the learners, giving them the opportunity to reflect on their previous experiences, as this teacher explained:

Wir haben jetzt noch keinen kompletten Kurs abgeschlossen, wir freuen uns, dass jetzt grade ein Vorbereitungskurs auf die digitale Prüfung bei uns im Bereich DaF läuft. Und wir haben bei der Entwicklung der Inhalte darauf geachtet, dass auf jeden Fall Raum und Zeit ist für diese Bewusstmachung und auch für den Austausch gibt. Immer wieder der Austausch mit anderen Studierenden, welche Erfahrungen habt ihr denn gemacht, wo steht ihr an der Stelle? Um auch da was mitzunehmen. Ich glaube, dass dieser Aspekt sehr wichtig ist./We haven't yet completed a full test preparation course, but we are happy that there is one preparation class for the digital TestDaF running right now. In developing the course and selecting the material, we paid attention to having enough time for these awareness-raising activities and for discussions within the group. The exchange with other learners on where they stand, to take something away for oneself. I think this is an important aspect.

That these activities were not only considered useful by the teachers, but may also have been perceived as helpful by their students, is suggested by the experience of one teacher who used these activities in relation to listening:

Ich habe diese Aufgaben auf der Lernplattform angelegt und wir haben dann während der Online-Phase die Situationen gesammelt, das haben wir dann in der nächsten Sitzung besprochen. Ich glaube, das hat gutgetan.

Im Modul „Hören“ gibt es viele Aufgaben und man hat Angst. Es hat sehr geholfen, dass die Kursteilnehmenden gesehen haben, dass sie nichts Außerirdisches, also nichts sehr, sehr Ungewöhnliches machen müssen, nur das, was sie auch aus ihrem Schulalltag kennen, sie müssen das jetzt auf Deutsch in der Prüfung nochmal machen. Ich finde, das war in der psychologischen Hinsicht sehr nützlich. Also hat beruhigt, mindestens./*I put these awareness-raising activities on our learning platform, and we discussed the situations [in which learners usually listen to a text] in our online meeting. I think that helped. The listening component [in the digital TestDaF] includes many tasks and some learners get scared. It helped that the learners could see that this is nothing from outer space, nothing unusual that they have to do in the tasks, only something they already knew from their every day school life. And now they just have to do it again in German. I think that this was useful from a psychological point of view, at least reassuring.*

These excerpts show that teachers were keen to address learners' preconceptions and concerns about the test. They clearly appreciated the opportunities of the awareness-raising tasks to familiarise learners with task requirements and their own skills in relation to these requirements. Even though the learning-oriented approach in general and the awareness-raising activities in particular were perceived as useful from a teacher's perspective, teachers also raised concerns about how learners would react to this new kind of test preparation, as one teacher put it:

Dieses Nachdenken über sich selbst und die eigene Vorgehensweise, die ist in großen Teilen unserer Teilnehmerschar nicht besonders ausgeprägt./*This self-reflection about approaching a task, this is something that is not very pronounced among our participants.*

Teachers suggested different reasons for this possible reticence. For one thing, they were sceptical about how learners from different cultural backgrounds would be able to adapt to these new learning strategies, as the following excerpt shows:

Aber die Frage war, wie können wir die Ziele dieses Konzeptes auf der Ebene der einzelnen Aufgabenstellungen so verständlich formulieren, ohne vorauszusetzen, dass die Teilnehmenden im Vorfeld bis jetzt als reflexionserfahren zu uns kommen. Weil manche kommen vielleicht aus Kulturen, dass sie sowas zum ersten Mal hören oder aufgrund ihrer kulturellen oder bildungskulturellen Prägung diesen ersten Schritt erstmal gar nicht zulassen, weil es unhöflich ist, zum Beispiel./*But the question was, how can we make the aims of the whole approach transparent on the level of the single learning activities, without taking it for granted. Because some of them come from cultures where this is not common, or their cultural or educational background doesn't allow for this first step [i.e. asking for prior experience], because it is regarded as rude, for example.*

In the interviews teachers differentiated between different groups of learners and their reaction to the learning-oriented approach. Some teachers recognised differences between learners at different proficiency levels:

Grade die B2-Kandidaten sind halt doch eher: Ich will die Prüfung machen und dann mal gucken. Was ich dann für das Studium brauche oder für die Ausbildung oder was, daran denke ich nicht. Mir geht es um die Prüfung. [...] Aber es passt halt nicht, diese Einstellung, zum digitalen TestDaF und da tun sich manche Teilnehmer – nicht alle, aber manche – dann schon schwer zu akzeptieren, ok, es geht nicht um die eine Aufgabe, es geht nicht um den einen Test, sondern es geht um das generelle Lernen von Kompetenzen. */Especially the B2 learners are more like: I want to pass the test, and then I'll see. I don't think about what I need for studying or other education like vocational training. I only care about the test. [...] But this attitude doesn't work for the digital TestDaF, and some participants – not all, but some – have difficulties to accept that it isn't about the ONE task, or about the ONE exam, it's about learning underlying competencies.*

According to them, learners at a lower B2 level were mainly interested in test-taking skills and single task requirements, while other teachers thought that the appreciation of the learning-oriented teaching material would be related to individual differences like, for example, a more general interest in language learning:

Es gibt immer einzelne Gruppen, einzelne Teilnehmende, die dafür sehr dankbar sind, solche Strategien zu lernen, um mit jedwedem Text zu jedwedem Thema zurechtzukommen, und es gibt immer Teilnehmende, die einfach wissen wollen, „so, jetzt hab ich dreimal A, dann muss ja wohl B drankommen“, also die test-wisen Studierenden. */There are always single groups, single test-takers who are more than grateful for learning these strategies to cope with each and every text on different topics; and there are test-takers who only want to know “so, now I have chosen three times A [as an answer], then it has to be B now” – these are test-wise students.*

Overall, teachers were aware that this new approach to test preparation would not only be a challenge for themselves, but also for the learners. They felt that their main task would be to convince learners of the benefits of this approach – the biggest one being the close link of test preparation and study preparation activities. The teachers' observations and reflections underscore the importance of identifying and addressing the specific LAL needs of teachers and learners in test preparation courses. For teachers, a good knowledge and understanding of the test construct seems essential in order to

offer appropriate and successful test preparation activities. For prospective test-takers it is important to understand the link between test tasks and TLU in the relevant domain. The data from the interviews has given evidence that test preparation materials based on a learning-oriented approach can be a helpful tool for teachers and learners to gain such knowledge and skills relating to language assessment.

Conclusion

The overall aim of this exploratory study was to evaluate perceptions of a learning-oriented approach for test preparation in the context of the digital TestDaF from a teacher and a learner perspective. In spite of the limited access to data, the interviews revealed some interesting insights into the need for teachers' LAL in the context of test preparation.

All teachers were very aware of the requirements of the TLU domain – not surprisingly since all of them had extensive experience in preparing international students for studying at institutions of HE in Germany. Regarding the test tasks of the digital TestDaF, teachers found them to be closely related to communicative tasks in academia. Task authenticity was mentioned as one of the most important aspects of the new test, and teachers hoped for a positive washback on test preparation classes since preparation activities would be similar to what is expected at university (Shohamy 1999, Wall and Horák 2011). It seemed important to them to make test-takers aware of the fact that the digital TestDaF promotes a closer link between preparing for the test and preparing for the requirements of the TLU. We would therefore argue that in test preparation classes, teachers need a profound understanding of knowledge of theory, as well as principles and concepts of language tests. Only by understanding the construct underlying the test tasks, the task requirements and how these are linked to the TLU domain, are teachers able to convey this to their students, thus enhancing their LAL as well.

The learning-oriented approach for test preparation and the developed materials were perceived as helpful. Teachers appreciated the accompanying teacher's guide that helped them in gaining a better understanding of the concept, focusing on the underlying competencies rather than on single task requirements. They especially stressed the importance of the awareness-raising activities that helped learners to build on their previous experience, hence focusing on their strengths and reflecting on useful strategies to cope with the requirements of the new tasks. Actively involving students 'so that they develop a better understanding of learning goals' (Carless 2007:59) therefore not only seems relevant in general language classes, but also crucial in the context of test preparation. Even if not always expressed explicitly, one thing the teachers aimed for was to enhance their learners' LAL. They also reported different aspects that they felt hinder learners'

LAL development in their language classroom: for example, limited time resources might restrict the opportunities to fully engage students in awareness-raising activities. Furthermore, many test-takers were mainly focusing on passing the test, and hence preferred the traditional approaches for test preparation like teaching-to-the-test. Teachers listed their students' cultural background and past learning experiences as well as different proficiency levels as possible reasons for this attitude towards assessment. In the future, it may therefore be helpful to gain more in-depth insights into learners' understanding of language assessment and test preparation. This study was able to address learners' perspectives through the eyes of experienced language teachers who were closely familiar with the target group of TestDaF. Still, interviews with prospective test-takers may offer new data that could be useful for creating opportunities for LAL development in language learners.

References

- Arras, U (2012) Im Rahmen eines Hochschulstudiums in Deutschland erforderliche sprachliche Kompetenzen: Ergebnisse einer empirischen Bedarfsanalyse, in Tinnefeld, T (Ed), *Hochschulischer Fremdsprachenunterricht: Anforderungen – Ausrichtung – Spezifik*, Saarbrücken: htw saar, 137–148.
- 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.
- Carless, D (2007) Learning-oriented assessment: Conceptual bases and practical implications, *Innovations in Education and Teaching International* 44 (1), 57–66.
- Cumming, A (2013) Assessing integrated writing tasks for academic purposes: Promises and perils, *Language Assessment Quarterly* 10 (1), 1–8.
- Davies, A (2008) Textbook trends in teaching language testing, *Language Testing* 25 (3), 327–347.
- Dresing, T and Pehl, T (2018) *Praxisbuch Interview, Transkription & Analyse: Anleitungen und Regelsysteme für qualitative Forschende (8. Aufl.)*, Marburg: Eigenverlag, available online: <https://www.audiotranskription.de/downloads/>
- Fulcher, G (2012) Assessment literacy for the language classroom, *Language Assessment Quarterly* 9 (2), 113–132.
- Gebril, A (2018) Test preparation in the accountability era: Toward a learning-oriented approach, *TESOL Journal* 9 (1), 4–16.
- Gesellschaft für akademische Studienvorbereitung und Testentwicklung (g.a.s.t.) (2019) *Konzept zur Vorbereitung auf den digitalen TestDaF*, unpublished manuscript.
- Gesellschaft für akademische Studienvorbereitung und Testentwicklung (g.a.s.t.) (2020) *Daten kompakt – Compact data*, available online: www.testdaf.de
- Green, A (2017) Learning-oriented language test preparation materials: A contradiction in terms?, *Papers in Language Testing and Assessment* 6 (1), 112–132.

- Hochschulrektorenkonferenz (HRK) (2021) *Hochschulen in Zahlen 2021*, available online: www.hrk.de/fileadmin/redaktion/hrk/02-Dokumente/02-06-Hochschulsystem/Statistik/2021-09-15_ONLINE_HRK-Statistikfaltblatt_2021_-_final.pdf
- Hochschulrektorenkonferenz (HRK) und Kultusministerkonferenz (KMK) (2020) *Rahmenordnung über Deutsche Sprachprüfungen für das Studium an deutschen Hochschulen (RO-DT)*, available online: www.hrk.de/fileadmin/redaktion/hrk/02-Dokumente/02-07-Internationales/Rahmenordnung_ueber_Deutsche_Sprachpruefungen_fuer_das_Studium_an_deutschen_Hochschulen_RO-DT_2020.pdf
- Inbar-Lourie, O (2013) Guest Editorial to the special issue on language assessment literacy, *Language Testing* 30 (3), 301–307.
- Jones, N and Saville, N (2016) *Learning Oriented Assessment. A systemic approach*, Studies in Language Testing volume 45, Cambridge: UCLES/ Cambridge University Press.
- Kecker, G and Eckes, T (2022) Der digitale TestDaF: Aufbruch in neue Dimensionen des Sprachtestens, *Informationen Deutsch als Fremdsprache* 49 (4), 289–324.
- Kecker, G, Zimmermann, S and Eckes, T (2022) Der Weg zum digitalen TestDaF: Konzeption, Entwicklung und Validierung, in Gretsch, P and Wulff, N (Eds), *Deutsch als Zweit- und Fremdsprache in Schule und Beruf*, Paderborn: Brill Schöningh, 393–410.
- Knoch, U, Huisman, A, Elder, C, Kong, X and McKenna, A (2020) Drawing on repeat test takers to study test preparation practices and their links to score gains, *Language Testing* 37 (4), 550–172.
- Malone, M E (2013) The essentials of assessment literacy: Contrasts between testers and users, *Language Testing* 30 (3), 329–344.
- Marks, D (2015) Prüfen sprachlicher Kompetenzen internationaler Studienanfänger an deutschen Hochschulen: Was leistet der TestDaF?, *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 20 (1), 21–39.
- Messick, S (1982) Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice, *Educational Psychologist* 17, 67–91.
- Messick, S (1996) Validity and washback in language testing, *Language Testing* 13, 241–256.
- Norris, J and Drackert, A (2018) Test review: TestDaF, *Language Testing* 35, 149–157.
- O’Sullivan, B, Dunn, K and Berry, V (2021) Test preparation: an international comparison of test takers’ preferences, *Assessment in Education: Principles, Policy & Practice* 28 (1), 13–36.
- Plakans, L (2013) Assessment of integrated skills, in Chapelle, C (Ed) *The Encyclopedia of Applied Linguistics*, Hoboken: Wiley.
- Popham, W (2009) Assessment literacy for teachers: Faddish or fundamental?, *Theory Into Practice* 48 (1), 4–11.
- Saldaña, J (2016) *The Coding Manual for Qualitative Researchers*, Thousand Oaks: SAGE.
- Shohamy, E (1999) Language testing: Impact, in Spolsky, B (Ed), *Concise Encyclopedia of Educational Linguistics*, Oxford: Pergamon, 711–714.
- 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.
- Wall, D and Horák, T (2011) *The impact of changes in the TOEFL exam on teaching in a sample of countries in Europe: Phase 3 and 4*, TOEFL Research Report RR-11-41, Princeton: Educational Testing Service.
- Zimmermann, S (2009) Falsche Vorbereitung? Erkenntnisse aus Teilnehmerleistungen der Prüfungsteile Mündlicher und Schriftlicher Ausdruck im Test Deutsch als Fremdsprache (TestDaF), in Yu, X (Ed) *TestDaF-Training und Studienvorbereitung: Beiträge zur chinesisches-deutschen Fachkonferenz: „TestDaF-Training und Studienvorbereitung“ vom 11. bis 12. Oktober 2008 am Deutschkolleg der Tongji-Universität Shanghai*, München: Iudicium, 63–83.
- Zimmermann, S (2021) *Test-takers' perception of the digital TestDaF*, paper presented at the BAAL TEA SIG Online-Conference, 17 September.

7 An online self-study course for Celpe-Bras raters and examiners: Induction into assessment practice and reflection

Margarete Schlatter

*Universidade Federal do Rio Grande do Sul (UFRGS),
Brasil*

Matilde Virginia Ricardi Scaramucci

Universidade Estadual de Campinas (Unicamp), Brasil

Juliana Roquete Schoffen

*Universidade Federal do Rio Grande do Sul (UFRGS),
Brasil*

This chapter analyses the design of an online self-study course for the raters and examiners of the Certificate of Proficiency in Portuguese for Foreigners (Celpe-Bras). It focuses on:

- The essential features of Celpe-Bras
- The theoretical framework of the course design based on induction into assessment practice and reflection
- How practice and feedback are integrated to promote the development of knowledge and skills in language assessment in an online self-study modality

Este capítulo analisa o desenho de um curso online autoformativo para avaliadores e examinadores do Certificado de Proficiência em Português para Estrangeiros (Celpe-Bras). A análise focaliza:

- As características fundamentais do Celpe-Bras
- Os pressupostos teóricos do desenho do curso baseado na indução à prática avaliativa e à reflexão
- De que modo a integração entre prática e feedback pode promover a construção de conhecimento e de habilidades para avaliar proficiência linguística em uma modalidade online autoformativa

Introduction

The Certificate of Proficiency in Portuguese for Foreigners (Celpe-Bras), developed and administered by the Brazilian Ministry of Education, is a large-scale task-based performance test developed to assess proficiency in listening, speaking, reading and writing. Launched in 1998, the exam is administered today in 126 accredited centres (48 in Brazil and 78 in other countries) (Inep 2020). Due to its construct built on the notions of genre and language use, its uses in academic and professional contexts, and the consistency accomplished in the rating procedures, Celpe-Bras has had considerable positive washback on the teaching and learning of Portuguese, and has been contributing to an increase in levels of assessment literacy among examiners, raters and users (Scaramucci 2008, 2016).

According to Davis (2022:322), in language performance assessment ‘raters perform the key role of transforming evidence of language ability into a score. Raters therefore have considerable impact on the extent to which scores reflect the intended construct, are reliable, and are fair for different types of individuals’. The procedures that have been used to align a rater’s decision-making process to the exam construct and to other raters (inter-rater reliability) include the use of scoring rubrics, the analysis of exemplar performances, and examiners’ training (Fulcher 2003, Weigle 2002). In this chapter, we present the design of a recently developed two-module online self-study course administered to Celpe-Bras raters and oral examiners. We then discuss the strategies, such as the practice–reflection–practice cycle, used to enhance the participants’ understanding of the underlying concepts and assessment procedures of the exam and, simultaneously, raise their awareness of their professional and social responsibilities.¹

Celpe-Bras: Assessing the use of Portuguese in social practices

Through a single exam, Celpe-Bras assesses six levels of proficiency and certifies four levels (Intermediate, High Intermediate, Advanced, and High Advanced). The exam is divided into two parts, a written part (three hours) and an oral part (20 minutes). Based on the construct of language use – ‘language is used for doing things’ (Clark 1996:3) – examinees are expected to use Portuguese in varied social practices that require reading, writing,

¹ Commissioned by Instituto Nacional de Estudos e Pesquisas Anísio Teixeira (Inep) with the collaboration of Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos (Cebraspe), both course modules were designed by the authors of this chapter, and developed and implemented under their supervision, with the collaboration of Cirlene S. Sanson, Ellen Y Nagasawa, Gabrielle R Sirianni, Giovana L Segat, Kaiane Mendel, Kétina A S Timboni, Leticia S Bortolini, Moisés Sousa, and Patricia M C Almeida.

listening, and speaking. The use of a single exam to assess distinct levels of proficiency presupposes that language users oftentimes need to participate in a variety of social practices regardless of how experienced they are. It is the management of their participation in the varied complex contexts that signals the extent of their proficiency in mobilising repertoires of cultural, discursive, and linguistic resources to do things with language more (or less) adequately, considering the sociolinguistic demands of the practices they need to or wish to engage in.

The concept of language use is transposed to the written part of the exam in the form of two listening-to-write and two reading-to-write tasks, each of which sets out the genre of the text to be written, its purpose, and the interlocutors of the text. These elements, in turn, set up listening and reading purposes, thus creating a situated writing frame to guide the examinee's perspective to read/watch/listen and to write, aimed at taking part in an institutional (not private) discursive situation. The input texts supply the necessary information to build concrete and consistent support for writing the texts and to level the examinees as to their background knowledge required on the topic. Figure 1 shows an example of a reading-to-write task (with accompanying translation in English) in which the examinee is asked to write a letter after reading a newspaper article.

As can be seen in the example, the task writing frame comprises an enunciator (the president of the Association of Shop Owners in a city where there are no parklets), interlocutors (the association members), a text genre (a letter) and a writing purpose (to suggest investing in the installation of parklets by explaining what they are and justifying the investment).² The reading purpose is to identify information in the text that can be used to construct the arguments to convince the interlocutors of the benefits of investing in that specific project. Examinees are expected to retextualise information that is relevant for the construction of a new discursive situation, taking into consideration all the elements that make up that specific language use. The same writing frame is used to define the parameters to assess the discursive and linguistic quality of the texts. Each of the four texts written by the examinee is assessed by two independent raters with a holistic rating scale. The scoring rubrics describe six levels of appropriateness of the texts in terms of the extent to which the interlocution and the purpose are addressed and the consistency of the informational, discursive, and linguistic resources used to accomplish the task (Inep 2020:73).

In the oral part, examinees participate in an interaction about personal interests and three current topics. The topics are chosen by the examiner (from a set of 20 different prompts) based on the examinees' profiles, drawn

2 For a thorough analysis of Celpe-Bras tasks, see Schoffen et al (2018).

Figure 1 Example of Celpe-Bras reading-to-write task



Certificado de Proficiência em Língua Portuguesa para Estrangeiros

Tarefa 3 | Parklets

Página 6

Você é o presidente da Associação de Comerciantes de uma cidade onde ainda não há *parklets*. Após ler a reportagem, decidiu sugerir que os associados invistam na instalação dessas estruturas. Escreva a eles uma carta, explicando o que são os *parklets* e justificando a importância do investimento.

Parklets ganham espaço e caem no gosto de Belo Horizonte



FOTO: Jair Amaral/EM/D.A. Press.

Em meio ao movimento acelerado de carros e ao vaivém de pedestres em áreas agitadas da capital, o belo-horizontino passou a ter novos espaços para relaxar, bater papo, ouvir música, ler, encontrar amigos, se divertir ou apreciar a paisagem. As varandas urbanas, como vêm sendo chamadas as minipraças do projeto BH *Parklets* – espaços de convivência rentes à calçada, em vagas de estacionamento –, já estão montadas em três pontos da capital, com previsão de novo endereço já na semana que vem.

Um dos responsáveis pela implantação da unidade é o empresário Alexandre Horta, sócio-proprietário do Deck Bar, em frente à varanda com estrutura de madeira, jardim, *wi-fi* e ponto de energia. A aposta no projeto, segundo ele, está atrelada à ideia de aumentar o uso público dos espaços urbanos.

“Nos interessamos já na época do lançamento do edital, porque vimos que poderíamos ajudar a arejar esta região”, afirma. Segundo ele, o investimento para instalação da estrutura foi de R\$ 20 mil. “Nosso objetivo não é ter retorno financeiro, mas contribuir para o bem comum da cidade”, disse.

Na prática, o projeto BH *Parklets* funciona assim: a prefeitura licencia a estrutura a pedido de representantes da iniciativa privada, que são responsáveis pelo custeio da instalação. De acordo com o idealizador do projeto, Luamã Lacerda, mesmo com o financiamento e a divulgação da marca, não há vínculo comercial, ou seja, as pessoas que usam o espaço não têm obrigatoriedade de consumir no estabelecimento. “Não há restrição de atendimento a quem estiver sentado e queira consumir. Mas qualquer pessoa pode usufruir

do espaço, a qualquer hora, porque o local é público”, explica Luamã. Segundo ele, as varandas urbanas já existem em São Paulo, Fortaleza e Porto Alegre.

No centro da cidade, onde a varanda foi montada em uma rua de grande circulação de veículos e pessoas, o projeto já ganhou adeptos fiéis. Moradores da via há 15 anos, o casal de aposentados Lúcia Sartori Sena e José Sena Reis comemora a implantação do espaço em um ponto tão movimentado da cidade. “Em todo esse tempo, vimos algumas mudanças positivas, como a retirada dos camelôs e a redução dos assaltos. Mas esse projeto é uma das melhores iniciativas, pois oferece entretenimento para quem mora na região”, afirma José.

Dono da loja Ortobom, que banca a instalação, o empresário e vice-presidente de Educação da Câmara de Dirigentes Lojistas de Belo Horizonte (CDL-BH), Marcos Ineco, fala do sucesso do projeto. “As pessoas entram na loja para elogiar e agradecer. O resultado tem sido muito positivo”, disse.

Disponível em: <http://www.em.com.br/app/noticia/gerais/2015/07/26/interna_gerais.672268/parklets-ganham-espaço-e-caem-no-gosto-de-belo-horizonte.shtml>. Acesso em: 7 jul. 2016 (adaptado).

2019–1 Celpe-Bras written part – Task 3 (Inep 2020:37).

Task 3 – Parklets

You are the president of the Association of Shop Owners in a city where there are no parklets yet. After reading the newspaper article, you decided to suggest to the association members that they should invest in the installation of such urban spaces. Write a letter to them explaining what parklets are and justifying the importance of the investment.

Synthesis of the text:

Parklets gain space and are appreciated in Belo Horizonte

The text talks about the recent installation of parklets in Belo Horizonte and the positive reaction of the community to these urban places. Aiming at increasing public use of urban spaces, parklets are installed alongside the sidewalks, in former kerbside parking spaces, and give the population new spaces to relax, meet friends, listen to music, have fun, or enjoy the scenery. Parklets are licensed by the City Hall at the request of shop owners, who are responsible for the cost of the installation. The shop can attach their brand to the parklet, but people who use the space are not required to consume at the establishment. According to a shop owner responsible for one of the parklets, the goal of the project is not to have financial return, but to contribute to the welfare of the population. Citizens appreciate the initiative, as it offers community space for those who live nearby. According to the Vice-President of the Association of Shop Owners in Belo Horizonte, the results have been very positive. In Brazil, besides Belo Horizonte, parklets have also been installed in São Paulo, Fortaleza and Porto Alegre.

from a questionnaire they answer during registration. Examiners are instructed to conduct an interaction starting by asking examinees to expand some aspects of their answers to the questionnaire and then to discuss the three different topics (the selected prompts), motivating them to develop their ideas by justifying, illustrating, comparing pros and cons, etc. Oral proficiency is assessed by two independent examiners: the interlocutor, who uses a holistic six-level-rubric rating scale, and the observer, who uses an analytical scale encompassing comprehension, interactional competence, fluency, lexical adequacy, grammatical adequacy, and pronunciation (Inep 2020:75).

Celpe-Bras online self-study course for examiners and raters

The oral examiners and text raters are Portuguese teachers who were required to take specific in-person training to learn about the rating procedures with hands-on assessment activities. Due to the COVID-19 pandemic in 2020, the training had to be developed fully online, which required a self-study course

followed by synchronous activities. A 25-hour module was first offered in 2020 to 492 oral examiners, while a second updated edition was offered in 2021 to 858 examiners. An 18-hour module for text raters was first trialled in 2022 with 28 raters.

In this section, an overview of the contents of each module is presented, followed by a discussion of some tasks and feedback to illustrate the design intended to promote practice–reflection–practice. The theoretical and practical content for both modules is presented through various interactive tasks, intended to provide experiences for assessing oral performance, conducting oral interactions, rating written texts and reflecting on assessment actions and attitudes. The content not only offers participants the opportunity to deepen their understanding of the exam construct, assessment criteria and procedures, but also to practise the use of the holistic and analytic rating scales with a sample of past exam recordings and texts. In both modules, recordings and texts of past tests are also used to illustrate the six proficiency levels of the exam and to practise making decisions based on scoring rubrics and performance variation. In the module for oral examiners, recordings are also used to illustrate and practise good rater behaviour and attitudes, as well as to guide and reflect on how to conduct the interactions and avoid bias. Table 1 presents the contents of both modules: five units for oral examiners, and four for text raters.

A self-study course presupposes that the participants can follow it without the presence of a mediator. Raters and oral examiners that will be part of the assessment team in the forthcoming Celpe-Bras edition are enrolled in their respective modules and, as soon as they have access to the online learning platform, can organise their agenda to develop the course tasks autonomously within a period of approximately 30 days. Participants have the help of technical support, but there are no course moderators. Learning mediation takes place through automated feedback (explanations and reflections) based on the participants' responses to the tasks (the participant's choice among alternative responses to the task triggers different feedback).

The course was designed for experienced and non-experienced participants: a diagnostic test at the beginning of the modules indicates whether the participant should study all the contents or just a selection of them. A final test shows the extent of the participant's learning and suggests further studies. The *diagnostic unit* is a prerequisite for accessing all the others: based on the results of the diagnostic test, participants receive tailored automated recommendations on the most relevant contents and are invited to set their own learning goals by selecting, from a list of assessment challenges and course objectives, the ones they wish to study. In the *practice units*, participants may choose whether to follow the proposed order of tasks or to follow their own pathway through the material considering their goals, time available and motivation. Aiming to promote the participants' learning

Table 1 Contents of the self-study course for Celpe-Bras examiners and raters

Self-study module for oral examiners	Self-study module for text raters
Introduction (in each module)	
Objectives and methodology Course materials and information regarding ethical procedures (recording/text de-identification) Navigation and participation guidelines Assessment of learning Course contents Pass requirements and performance rating scales Timetable	
U1 – Welcome, setting learning goals, diagnostic test	U1 – Welcome, setting learning goals, diagnostic test
<ul style="list-style-type: none"> • Welcome video: Reasons to take the course, course objectives and contents • Diagnostic test: One task about U2 contents (oral interaction procedures and conduction); two tasks about U3 contents (holistic and analytic assessment of oral performances); one quiz about U4 contents (exam construct) • Setting learning goals 	<ul style="list-style-type: none"> • Welcome video: Reasons to take the course, course objectives and contents • Diagnostic test: One task about U2 contents (holistic assessment of texts); one quiz about U3 contents (exam construct) • Setting learning goals
U2 – Conducting the oral interaction	U2 – The examinee’s assessment (scoring texts)
<ul style="list-style-type: none"> • Four tasks, namely: the specifications of the exam’s oral part; procedures to prepare for the conduction of the interaction; procedures to conduct the interaction; evaluating the appropriateness of the examiner’s procedures to conduct the interaction 	<ul style="list-style-type: none"> • One task to analyse the rating scales in detail, identifying what distinguishes the six levels of proficiency • Three tasks to use the rating scales to assess examiners’ performances focusing on different assessment challenges
U3 – The examinee’s assessment (scoring oral performance)	
<ul style="list-style-type: none"> • One task to analyse the rating scales in detail, identifying what distinguishes the six levels of proficiency • Two tasks to use the rating scales (holistic and analytic) to assess examiners’ performances 	
U4 – The relation between interaction procedures, the exam construct and the specifications of the exam’s oral part	U3 – The relation between rating procedures and the exam construct
<ul style="list-style-type: none"> • Four tasks to relate the exam’s oral part procedures, the holistic and analytic rating scales and the interaction characteristics to the exam specifications and theoretical construct • One task to synthesise the profile of the oral examiner 	<ul style="list-style-type: none"> • Five tasks to relate the exam’s written part rating procedures and the holistic rating scales to the exam specifications and theoretical construct • One task to synthesise the profile of the text rater

Table 1 (continued)

Self-study module for oral examiners	Self-study module for text raters
U5 – Assessment of learning <ul style="list-style-type: none"> • Two tasks to score oral performances (using holistic and analytic rating scales) and to justify decisions • One task to evaluate and justify the appropriateness of oral interaction procedures and conduction of interaction • One quiz about the exam construct 	U4 – Assessment of learning <ul style="list-style-type: none"> • One task to score a set of texts (using the holistic rating scale) and to justify decisions • One quiz about the exam construct
Conclusion (in each module) Self-assessment and evaluation of the module	

autonomy, tasks in these units have automated correction and feedback, and a self-assessment at the end. They may also choose whether to develop all tasks (and all items in a task), or skip or repeat some of them. In the *assessment unit*, participants are invited to score new sets of oral/written performances and answer a quiz on the exam construct. The pass grade is 70%; below that, participants are guided to go deeper into the practice units to prepare for a second attempt, this time on a set of oral/written performances and questions that combine repeated and new recordings/texts and items.³

In line with the self-study nature of the course, the participant receives indication of success or failure after concluding each one of the tasks, as well as detailed feedback presented in videos, audios, onscreen texts, or closing tasks requesting a synthesis of what was studied. Traditionally, online self-study courses follow a presentation-assessment approach, with tasks that require participants to study contents (read a text or watch a video) to answer questions (or relate columns, drag and drop, etc.), to which automatic correct/incorrect feedback is provided. In contrast with this transmission model of learning and drawing on the work of Dewey, Piaget, Vygotsky, and Freire, who advocate experience and reflection on doing as key to learning, the design of the Celpe-Bras modules proposes induction into assessment practice and reflection in a practice–reflection–practice cycle of tasks. According to Kolb and Kolb (2013:6–7), *experiential learning* rests on six assumptions:

1. Learning is a process that derives from experience and does not end at an outcome.
2. Learning is always re-learning.

³ The interactive content of the practice units was implemented in H5P, a platform which allows deliberate and recursive practice without recording or assessment. The diagnostic test and the learning assessment tasks were implemented in Moodle so that the results can be recorded and assessed.

3. Learning requires the resolution of conflicts between dialectically opposed demands to adapt to a new situation.
4. Learning is a holistic and integrative process of adapting to the world, involving thinking, feeling, perceiving, behaving, problem solving, decision making and creativity.
5. Learning involves one's synergetic interplay with the environment.
6. Learning creates knowledge as the result of the constant transaction between co-constructed social knowledge and personal knowledge.

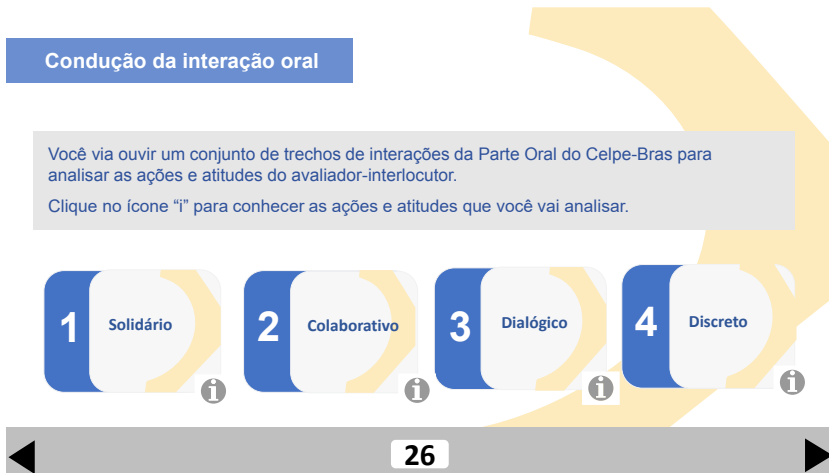
The course tasks were designed to create meaningful experience for examiners, i.e., 'conflicts' to be resolved which demand mobilising existing resources, beliefs and ideas about assessing oral and written performances, 'so that they can be examined, tested and integrated with new, more refined ideas' (Kolb and Kolb 2013:7). The new perspective is presented via feedback, which creates opportunities for participants to 'move between opposing modes of reflection and action and feeling and thinking', thus recreating knowledge (Kolb and Kolb 2013:7).

As illustrated by the tasks below, participants are requested to face assessment challenges that demand the integration of their prior experience, abstraction, reflection, and action. Some assessment issues that cause conflicting modes of resolution and that are included in the course are: *Are there more/less salient oral/written criteria at distinct levels of proficiency or for users of close/distant languages? To what extent can the actions and attitudes of the oral examiner interfere in the performance of the examinee? How does knowledge of the theoretical construct of the exam help understand assessment criteria and assess performance?* Answers to these questions are not straightforward once they draw on prior experiences, skills, opinions and attitudes regarding assessment practices, and demand both a situated analysis of each written or oral performance and also knowledge of the exam's expectations in relation to assessment procedures and the use of scoring rubrics.

To deal with this complexity, both modules of the course acknowledge a range of more-to-less acceptable responses to the assessment challenges presented, explaining preferred responses and calling the participants' attention both to the required use of the guidelines and parameters and to possible adjustments considering the uniqueness of each performance. After solving each task, specific feedback invites participants to compare their answers with suggested solutions and justifications so that they can confront their perspectives relative to the Celpe-Bras construct. This comparison of solutions aims at creating opportunities for the examiners to confirm, adapt, re-learn, re-construct justifications for their actions and decisions, and develop expertise on the Celpe-Bras assessment standards. To illustrate the practice–reflection–practice cycle

as course design, Figure 2 shows three moments of a sequence of Unit 2 interactive screens in a task on the examiner's attitudes and actions while conducting the oral interaction.

Figure 2 Practice–reflection–practice as course design⁴



Conducting the oral interaction

You will hear a set of excerpts of interactions of the oral part to analyse the actions and attitudes of the interlocutor.

Click on the 'i' icon to learn about the actions and attitudes you will analyse.

1-Supportive 2-Collaborative 3-Dialogical 4-Discreet

The pop-ups offer brief definitions, such as:

Dialogical: co-constructs the topic of interaction, adapting the questions to what was said, maintaining the proposed topic, requesting the examinees to expand or deepen ideas, without intimidating them or putting them in an embarrassing situation.

⁴ The screens used in this chapter to illustrate course tasks belong to the oral part module, trialled in 2020 and revised in 2021. The written part module was trialled in 2022 and is currently undergoing adjustments based on the analysis of the results and perceptions of the participants.

An online self-study course for Celpe-Bras raters and examiners

Em cada interação, você vai avaliar uma ou mais ações e atitudes do avaliador-interlocutor como **adequadas**, **parcialmente adequadas** ou **inadequadas**.

Clique em cada interação para ser direcionado ao áudio, às perguntas e as materiais utilizados no trecho. Após concluir a análise da interação A, você será redirecionado automaticamente a esta tela inicial para poder selecionar a próxima interação.



INTERAÇÃO A	INTERAÇÃO B	INTERAÇÃO C	INTERAÇÃO D
INTERAÇÃO E	INTERAÇÃO F	INTERAÇÃO G	INTERAÇÃO H

27

In each interaction, you will analyse one or more actions and attitudes of the interlocutor as adequate, partially adequate or inadequate.

Click on each interaction to get access to the recording, questions, and materials used in the excerpt. After you complete the analysis of Interaction A, you will be automatically redirected to this home screen so that you can select the next interaction.

Como você pôde observar neste trecho, a avaliadora-interlocutora mostra uma atitude solidária e colaborativa durante a interação, acompanhando a fala do examinando através de sinalizações de anuência, concordância, empatia e exclamações, demonstrando seu interesse por aquilo que está sendo dito. Muito interessante a forma como a avaliadora usa feedbacks como forma de incentivo para que o examinando dê continuidade à sua fala. Particularmente ilustrativa de sua atitude dialógica é o trecho em que a avaliadora tenta auxiliar o examinando a entender o significado de "Lar doce lar", fundamental para o entendimento do EP utilizado, que tem como título "Celular doce lar". Essa atitude de co-construção do tópico, fazendo perguntas que pudessem levar o examinando a entender o tema a partir de suas respostas, foi importante para ele prosseguir, apesar de muitas dificuldades na expressão de suas ideias. Além disso, a avaliadora-interlocutora, em nenhum momento, teve uma atitude intimidatória ou que colocasse o examinando em uma situação constrangedora por não ter entendido o EP.

[Ouça ou leia o comentário sobre esta interação.](#)  

30

As you could observe in this excerpt, the interlocutor shows a supportive and collaborative attitude during the interaction, assenting, using exclamations and empathy expressions to show her interest in what is being said. Remarkably interesting is the way the examiner uses feedback as an incentive for the

examinee to continue his speech. Particularly illustrative of her dialogical attitude is the passage in which she helps the examinee understand the meaning of 'Home sweet home', crucial for the understanding of the prompt 'Cell phone sweet home'. This attitude of co-construction of the topic, asking questions that could lead the examinee to understand the topic, was important for him to continue, despite his many difficulties in expressing the ideas. Moreover, the examiner, at no time, showed an intimidating attitude or put the examinee in an embarrassing situation because he did not understand the prompt.

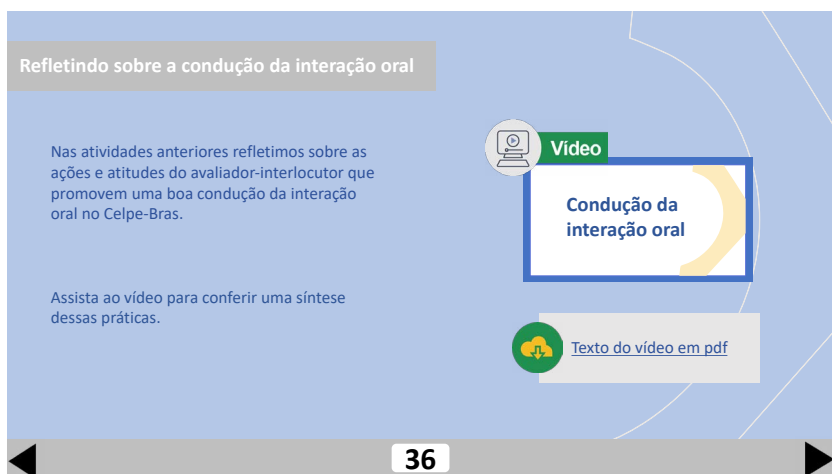
As can be seen above, after learning about the actions that define the four examiners' attitudes, participants are invited to analyse excerpts of interactions so that they can understand possible impacts of their actions on an examinee's performance. Participants can take charge of their own learning by deliberately choosing from various items (in this case, eight items, as shown on Screen 27 in Figure 2) and deciding on the extent of their practice (by choosing some, all or repeating them), which gives them space to monitor and control their attention, effort, and time on task. After completing each item, participants can compare their responses to the preferred one and read a justification that analyses that specific performance according to the exam's construct. In the task above, for instance, it is expected that participants learn what it means to be supportive, collaborative, dialogical and discreet not only by reading the definitions (Screen 26 in Figure 2) but by analysing the appropriateness of the examiner's actions in the excerpts and, inspired by the feedback, reflecting on and calibrating their prior knowledge to create new knowledge. This design presupposes that the development of expert performance requires deliberate practice through tasks that 'are initially outside their current realm of reliable performance, yet can be mastered within hours of practice by concentrating on critical aspects and by gradually refining performance through repetitions after feedback' (Ericsson 2006:694).

According to Clark and Mayer (2016:276), learners' answers offer *teachable moments*, and to make good use of these moments to promote learning, feedback should incorporate explanation. Having this in mind, course planning demanded: a) gathering examples of authentic assessment challenges; b) organising the contents (exam construct definitions and procedures) to explain the preferred solutions; c) constructing tasks that (re)create the challenging assessment situations with different possible solutions that may promote teachable moments; and d) writing explanatory feedback tailored both for preferred, less preferred and incorrect answers in a way that the participants' background knowledge is respected and built on. The feedback is both process feedback, which 'focuses on the strategies used to arrive at the response', and self-regulatory feedback, which 'directs the learners to monitor their response and reflect on their learning' (Clark and Mayer 2016:279). Moreover, as suggested by Clark and Mayer (2016:179),

feedback is given right after the participants' answers, on the same screen of the task and using 'a conversational style of writing or speaking (including using first- and second-person language), polite wording ... and a friendly human voice'.

To end and summarise the practice on attitudes and actions while conducting the oral interaction, participants are led to a concluding reflection on the topic and then to a link to additional guidelines (Screen 36 in Figure 3) on how to select and calibrate questions to examinees and how to deal with unexpected situations, such as a recording interruption or examinees' emotional or physical stress.

Figure 3 Practice–reflection–practice as course design: concluding feedback



Reflecting on the conduction of oral interaction

In the previous activities, we reflected on the actions and attitudes of the interlocutor that promote good conduction of oral interaction in Celpe-Bras.

Watch the video for a synthesis of these practices.

[Link to video/Link to video script]

Feedback in different modalities aims to motivate participants with different learning styles. Besides audio, video and onscreen texts, feedback is also offered in closing tasks in which participants organise and synthesise previously studied content in tables, lists or paragraphs; instead of reading only, participants actively use the text they are reading to construct assertions, relate the skills just practised to exam guidelines, or perform other actions to systematise knowledge. For example, the closing task in Figure 4 aims at systematising criteria for decision-making on proficiency levels after a sequence of oral performance assessment using the analytical scale.

Figure 4 Practice–reflection–practice as course design: closing tasks

Atividade - Critérios para a tomada de decisão sobre os níveis de proficiência

Complete as asserções abaixo com os trechos a seguir. Arraste os trechos para os espaços adequados.

Para decidir o nível de proficiência de compreensão oral, é necessário levar em conta	até que ponto o examinando é capaz de atribuir sentido ao fluxo natural da fala de seu interlocutor.	à frequência em que ocorrem limitações e inadequações e a seus impactos na interação, causando possíveis rupturas no desenvolvimento do tópico.
Na competência interacional, o uso de estratégias de comunicação está relacionado à ocorrência de impasses na interação.		à frequência em que ocorrem inadequações em relação às convenções da variedade falada pelo examinando e aos seus impactos na interação, causando possíveis rupturas no desenvolvimento do tópico.
Pausas e hesitações são características da fala. Elas podem ser relevantes para definir níveis de proficiência da fluência se		à frequência em que ocorrem inadequações em relação às convenções da linguagem falada e aos seus impactos na interação, causando possíveis rupturas no desenvolvimento do tópico.
Os recursos lexicais são analisados em sua amplitude e adequação. A decisão sobre o nível de proficiência deve levar em conta a extensão das contribuições feitas e está vinculado		causarem interrupções no desenvolvimento da interação. É importante considerar também as pausas e hesitações em relação à extensão das contribuições feitas, ou seja, se causam interrupções pouco ou muito frequentes.
Os recursos gramaticais são analisados em sua amplitude e adequação. A decisão sobre o nível de proficiência deve levar em conta a extensão das contribuições feitas e está vinculado		
A pronúncia é analisada em relação à produção de sons, ritmo e entonação. A decisão sobre o nível de proficiência deve levar em conta a extensão das contribuições feitas e está vinculado		caso não ocorram impasses, as estratégias não se tornam relevantes. Caso ocorram, é importante analisar que tipos de estratégias são utilizadas (mais elaboradas ou menos elaboradas).

● Verificar resposta

Criteria for decision-making on proficiency levels

Complete the following assertions. Drag the snippets to the proper spaces.

Example (first assertion): Column 1: To decide on the level of proficiency of oral comprehension, it is necessary to consider/Column 2 (snippet that completes the assertion): to what extent the examinee is able to attribute meaning to the interlocutor’s natural flow of speech.

As mentioned earlier, the practice–reflection–practice cycle also shapes the overall course path: the participants start by using their prior knowledge in a diagnostic test and decide on their learning goals based on their prior experience, perceived needs, and wishes to learn about certain topics. The practice units provide space for recursive action; testing their knowledge; reflecting; and adjusting concepts, attitudes and feelings related to their learning goals by offering them the opportunity to (re)analyse all recordings/texts and concepts they encountered in the diagnostic test – now reorganised into interactive tasks with feedback providing detailed analysis, explanation, criteria and guidelines on how to improve assessment practices. By doing the tasks in the practice units, participants can pursue deep experiential learning to develop expertise related to the Celpe-Bras assessment procedures and construct their identity as a member of the Celpe-Bras community of examiners. In other words, by recursively interacting with experienced participants’ solutions to Celpe-Bras assessment practices (feedback), newcomers can gradually move from legitimate peripheral participation (in the course practices) to full participation as certified examiners. According to Lave and Wenger (1991:29):

Learning viewed as situated activity has as its central defining characteristic a process that we call *legitimate peripheral participation*. By this we mean to draw attention to the point that learners inevitably participate in communities of practitioners and that the mastery of knowledge and skill requires newcomers to move toward full participation in the sociocultural practices of a community. “Legitimate peripheral participation” provides a way to speak about the relations between newcomers and old-timers, and about activities, identities, artifacts, and communities of knowledge and practice. It concerns the process by which newcomers become part of a community of practice.

This approach is consistent with what could be considered an expanded view of assessment training and, as such, more adequate if the aim is to promote assessment literacy. To contribute to raising the levels of assessment literacy of examiners and raters – who are also teachers, in this case – the practice–reflection–practice cycle intends to go beyond training the technical skills to conduct an oral interview and to use a scoring scale to rate the examinee’s oral performance. Through contextualised practice followed by reflection and theoretical knowledge, and by a new cycle of reflection and practice, participants are invited to invest in developing both their assessment skills and abilities and, above all, a sound understanding of the principles of assessment as social practice, with its various roles, functions and impacts on society.


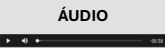





When participants taking the course consider they are ready, they can proceed to the tasks that will assess their learning (Figure 5). In this final unit, they will be asked to show a more informed practice by assessing new examinees’ recordings/texts, and justifying their score decisions. Being able to justify one’s actions presupposes deeper understanding once it requires the articulation of discursive knowledge that involves reflection and meta-analysis of the assessment procedures and learning experience. Having the opportunity to redo the assessment tasks – if they do not obtain a result above 70% – should help them (re)define their relationship to error as a relevant step to learning. It should also give them the possibility to (re)evaluate the role of practice–reflection–practice cycles in their own learning pathway as a process that involves experiencing, reflecting, thinking and acting, and of recursively confronting prior knowledge with other perspectives to (re)create knowledge.

In the learning assessment tasks, participants can test the skills required from Celpe-Bras examiners that they have practised before. According to Clark and Mayer (2016:275), skill-building demands ‘practice on the component skills that are required for a specific work domain’ and interactions that ‘require learners to respond in a job-realistic context’. This ‘mirror the job’ principle proposed by the authors is met both in the participants’ decision-making process proposed by the task and on the scores

Figure 5 Learning assessment task: using the holistic scoring scale

Questão 1
Ainda não respondida
Vale 3,00 ponto(s).

Interação A:
Ouça a interação na íntegra e marque uma nota para o desempenho do examinando. Clique nos ícones para ter acesso aos conteúdos que irão lhe orientar nessa avaliação.

 ÁUDIO 	 Grade Holística
 Questionário do Participante	 Roteiro 1
 Roteiro 2	 Roteiro 3

Escolha uma opção:

- 5 – Avançado Superior
- 4 – Avançado
- 3 – Intermediário Superior
- 2 – Intermediário
- 1 – Básico
- 0 - Elementar

Questão 2
Ainda não respondida
Vale 1,00 ponto(s).

Interação A:
Selecione a alternativa que melhor justifica a nota atribuída.

Escolha uma opção:

- Contribui muito para o desenvolvimento da interação, formulando respostas que expandem as ideias, com muita desenvoltura e autonomia. Compreende a fala em fluxo natural, salvo em raro momento, em que repete a pergunta como estratégia para sinalizar a falta de compreensão. Apresenta variedade ampla de vocabulário e de estruturas, com eventuais inadequações que produzem marcas de outras línguas, sem comprometimento da interação.
- Contribui para o desenvolvimento da interação, ainda que se limite, com frequência, a respostas breves. Demonstra dificuldade para manter o fluxo natural da fala, bem como alguns problemas de compreensão oral, o que pode levar a uma necessidade frequente de repetição/reestruturação por parte do interlocutor. Apresenta algumas inadequações de vocabulário, estruturas e/ou pronúncia que, em diferentes momentos, comprometem a interação.
- Contribui para o desenvolvimento da interação, não se limitando a respostas breves. Alguns problemas na compreensão do fluxo natural, com algumas necessidades de repetição. O fluxo da fala, apesar de algumas interrupções e de às vezes ser um pouco mais lento, apresenta certa naturalidade, principalmente ao falar sobre si, com menos fluência ao discutir os temas propostos. Apresenta alguns problemas de compreensão oral, além de algumas inadequações de vocabulário, estruturas e/ou pronúncia que produzem marcas de outras línguas e que, ocasionalmente, comprometem a interação.
- Contribui para o desenvolvimento da interação, formulando respostas que expandem as ideias, com desenvoltura e autonomia. Compreende a fala em fluxo natural, salvo em raro momento, em que repete a pergunta como estratégia para sinalizar a falta de compreensão. O fluxo da fala, em alguns momentos, pode ser um pouco mais reticente, mas em geral apresenta naturalidade ao falar de si e ao discutir os temas propostos. Apresenta variedade ampla de vocabulário e de estruturas, com poucas inadequações que produzem marcas de outras línguas, sem comprometimento da interação.

Interaction A: Listen to the interaction and score the examinee's performance. Click on the icons to get access to the contents needed for this assessment.

Contents: Recording/Holistic scale/Examinee's questionnaire/Prompts 1, 2 and 3

Choose one option:

5 – High Advanced/4 – Advanced/3 – High Intermediate/2 – Intermediate/1 – Basic/0 – Elementary

Interaction A:

Select the alternative that best justifies your score.

Example (first alternative): The examinee contributes a lot to the interaction, formulating responses which expand the ideas in a very articulated and autonomous way. Understands the natural flow of speech, except in rare occasions, in which echoing is used as a strategy to signal lack of comprehension. Uses a large and rich variety of vocabulary and grammatical structures; eventual inadequacies which show traces of other languages do not disrupt the interaction.

which are considered acceptable (100% to correct scores; 50% to scores that do not cause discrepancies; 0% to scores that cause discrepancy). As mentioned before, the discursive articulation of the justifications intends to call the participants' attention both to the required use of the parameters and to the uniqueness of each performance; the parameters should be interpreted so as to match the performance more closely with one of the proficiency levels described.

Participants are allowed to make a second attempt to score the performances. Davis (2022:330) explains that 'initial scoring practice may be followed by additional practice and review of scoring materials until the individual feels ready to attempt certification'. This opportunity of a second attempt to do the assessment tasks presupposes that, in a rater training that prioritises learning over teaching, participants should experience evaluation as learning, not as punishment. The instruction below shows the feedback offered to participants who score below 70%:

You presented inconsistencies in scoring the performances and/or in justifying your scores. We strongly recommend that you resume the contents of Unit 3. In this unit, you can learn or review the Celpe-Bras perspective on assessing oral proficiency and how to use the holistic rating scale. Your in-depth knowledge of the exam assessment parameters is fundamental for the validity and reliability of the exam results.

After reviewing the content on holistic assessment, you can try again with a new pair of interactions. You can choose to resume Unit 3 now and then make your second attempt or proceed to the next assessment task and resume this task later.

Considering that the course was designed to be one of the requisites to certify examiners and raters, both modules propose that participants become eligible for certification once they have scored a set of pre-scored responses with acceptable levels of accuracy (Davis 2022:330).

To further develop the participants' learning autonomy, at the end of each module they are invited to answer two questionnaires: a self-assessment

and an evaluation of the course. In the self-assessment task, they may reflect once again on the course objectives and answer to what extent they achieved their personal learning goals, considering their diagnostic test results, their learning path, and their overall achievements. The results of the self-assessment can further help them decide on new learning goals and whether they wish to study or review some course content. The course evaluation questionnaire requests that participants rate aspects of the online self-study modality, the relevance and effectiveness of tasks, and the alternatives they were given throughout the course to decide on their own learning paths, considering their diverse levels of previous experience and knowledge about the exam assessment practices.

Final remarks

According to Clark and Mayer (2016:265), ‘there is considerable evidence that well-developed practice interactions promote learning – especially in asynchronous e-learning’. We believe that the two-module course design discussed in this chapter, which draws on induction into assessment practice and reflection to develop language assessment literacy (LAL), has put into practice the five principles proposed by the authors to maximise the benefits of practice interactions (Clark and Mayer 2016:265):

- include sufficient practice to achieve the learning objective,
- require learners to respond in job-realistic ways,
- incorporate effective feedback to learner responses,
- distribute practice among the learning events,
- apply multimedia principles [to engage participants with different learning styles].

We also believe that the design has operationalised what Kolb and Kolb (2013:37) propose as teaching tailored to participants’ learning demands through opportunities to

... help learners organize and connect their reflection to the knowledge base of the subject matter ... provide models or theories for learners to use in subsequent analysis [provide] abstract conceptualization and active experimentation to help students apply knowledge toward performance goals ... monitor the quality of student performance toward the standards they set, and provide consistent feedback ... draw on concrete experience and active experimentation to help learners take action on personally meaningful goals.

With this analysis of the Celpe-Bras online course design, we intend to contribute to the understanding of how the development of knowledge and

skills in language assessment can be conceptualised and operationalised in an online self-study modality. The discussion may also provide a framework for future studies aiming at the analysis of Celpe-Bras and oral examiners' and text raters' performance in other exams, with implications both for increasing the validity and reliability of the exam, as well as for the enhancement of language professionals who wish to develop expertise in additional language teaching, learning and assessment practices.

References

- Clark, H H (1996) *Using Language*, Cambridge: Cambridge University Press.
- Clark, R C and Mayer, R E (2016) *E-learning and The Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning* (Fourth edition), Hoboken: Wiley.
- Davis, L (2022) Rater and interlocutor training, in Fulcher, G and Harding, L (Eds) *The Routledge Handbook of Language Testing* (Second edition), London: Routledge, 322–338.
- Ericsson, K A (2006) The influence of experience and deliberate practice on the development of superior expert performance, in Ericsson, K A, Charness, N, Feltovich, P J and Hoffman, R R (Eds) *The Cambridge Handbook of Expertise and Expert Performance*, Cambridge: Cambridge University Press, 685–705.
- Fulcher, G (2003) *Testing Second Language Speaking*, London: Longman.
- Instituto Nacional de Estudos e Pesquisas Anísio Teixeira (Inep) (2020) *Documento Base do exame Celpe-Bras*, Brasília: Inep, available online: www.ufrgs.br/acervocelpebras/wp-content/uploads/2021/12/Documento-base-do-exame-Celpe-Bras-2020.pdf
- Kolb, A Y and Kolb, D A (2013) *The Kolb learning style inventory – Version 4.0: A comprehensive guide to the theory, psychometrics, research on validity and educational applications*, Experience Based Learning Systems, available online: www.learningfromexperience.com
- Lave, J and Wenger, E (1991) *Situated Learning: Legitimate Peripheral Participation*, Cambridge: Cambridge University Press.
- Scaramucci, M V R (2008) O exame Celpe-Bras em contexto hispanofalante: percepções de professores e candidatos, in Wiedemann, L and Scaramucci, M V R (Orgs) *Português para Falantes de Espanhol: Ensino e aquisição – Portuguese for Spanish Speakers: teaching and acquisition*, Campinas: Pontes, 175–190.
- Scaramucci, M V R (2016) Letramento em avaliação (em contexto de línguas): contribuições para a Linguística Aplicada, Educação e Sociedade, in Jordão, C M (Org) *A Linguística Aplicada no Brasil: rumos e passagens*, Campinas: Pontes, 141–165.
- Schoffen, J R, Schlatter, M, Kunrath, S P, Nagasawa, E Y, Sirianni, G R, Mendel, K, Truyllo, L R and Divino, L S (2018) *Estudo descritivo das tarefas da Parte Escrita do exame Celpe-Bras: Edições de 1998 a 2017*, Porto Alegre: Instituto de Letras – UFRGS, available online: www.ufrgs.br/acervocelpebras/wp-content/uploads/2021/12/Schoffen-et-al-2018.pdf
- Weigle, S C (2002) *Assessing Writing*, Cambridge: Cambridge University Press.

8

Language assessment literacy in a workplace environment – an exploratory study from the testing of language proficiency in aviation

Neil Bullock

International Civil Aviation English Association

In this chapter I explore the role of language assessment literacy (LAL) within the system for assessing the language proficiency of air traffic controllers and pilots in aviation worldwide. I discuss the importance of diverse stakeholders' knowledge, skills and experience and how these may influence the effectiveness of this system, focusing on:

- Core features of the existing recommendations for assessment of language proficiency in aviation
- The expected knowledge and skills of the diverse stakeholders involved (e.g. pilots, air traffic controllers, language trainers, etc.)
- Some recent issues raised by key stakeholder groups concerning the effectiveness of test instruments and their link with assessment literacy
- Insights from recent research exploring the stakeholder groups' perceptions of the level of assessment literacy required for their roles
- Preliminary suggestions about how such insights might help to promote greater assessment literacy among stakeholders to address the issues raised

Introduction

Language assessment literacy in professional domains

Since the theories and concepts of *language assessment literacy* (LAL) began to emerge in the 1990s, a considerable amount of literature has been published on this topic. Such studies, however, have tended to focus on teachers and policy makers in traditional educational contexts, i.e. schools, colleges and universities (Fulcher 2012, Harding and Kremmel 2016).

Other issues focused on less extensively have included language assessment training (Inbar-Lourie 2008) and the need to address wider stakeholder interests (Taylor 2013).

Such research and debate are undoubtedly laudable in developing the concept of LAL, but arguably insufficient if we wish to support the growing importance of literacy across a wider range of assessment contexts, especially professional and vocational domains (Ingham and Thighe 2006, Taylor 2009). This need for further research is of particular importance where the real-world language use of test-takers in occupational settings diverges from a more traditional general language learning environment (Dudley-Evans and St John 1998, Kim 2018) and includes recognition that a more diverse group of stakeholders may be involved (Baker 2016, Inbar-Lourie 2013). Bachman and Palmer (1996) draw our attention to this specific workplace communication with the concept of Target Language Use (TLU), noting that test development must include a clear explanation of the purpose, test-taker and domain in which the language is used. A broader view of language assessment in professional settings was adopted by Jacoby and McNamara (1999:214) who advocated the concept of ‘indigenous’ criteria, focusing on the important elements involved during ‘naturally occurring [...] socialization [...] in professional settings’.

Such perspectives therefore suggest that a key principle in developing language assessment programmes for workplace domains is the identification of the skills and knowledge needed in each stakeholder role, although few studies have explicitly investigated this.

In this chapter I examine the system for assessment of language proficiency in the specific-purpose workplace domain of aviation, notably for pilots and air traffic controllers. In this chapter, the term *stakeholder* refers to anyone who, through their primary professional or semi-professional role(s), is involved in language assessment according to the International Civil Aviation Organization (ICAO) system of the Language Proficiency Requirements (LPRs). The term ‘language assessment literacy’ is conceptualised as the *perceptions, skills, knowledge* and *experience* required by stakeholders in carrying out their role within this system of the LPRs.

The chapter is composed of five parts. In the first, I present the core features of the existing system for assessing language proficiency among air traffic controllers and pilots in aviation. In the second, I go on to describe how the system attempts to view the expected knowledge and skills of the diverse stakeholders involved in the LPR system, and then in the third I highlight emerging issues and concerns voiced by certain stakeholder groups concerning the effectiveness of some test instruments and the possible links with assessment literacy. In the fourth part, I explain the stages of a recent research study which set out to identify the key stakeholders in this domain and provide evidence of the potential skills and knowledge (i.e. assessment

literacy) required for each stakeholder role, and then, in the fifth and final part, I explore how such insights might help promote greater assessment literacy among stakeholders and help address some of the issues raised.

Assessment of language proficiency in aviation

Existing recommendations for assessment of language proficiency in aviation

In 2008, ICAO implemented a worldwide system to assess proficiency of plain language¹ in speaking ability and listening comprehension for radiotelephone communication between pilots and air traffic controllers (ATCOs). The ICAO LPRs were developed and introduced following research showing that a lack of plain language English proficiency had been a causal factor in a series of major aviation incidents and accidents (Alderson 2009, International Civil Aviation Organization 2010). Whilst English is the particular focus of the LPRs, as it has been the designated common language for international aviation communications since the 1950s, most pilots and controllers worldwide must be tested for *any* language they use to communicate over the radiotelephone (Bullock 2015, International Civil Aviation Organization 2010).

ICAO, as an integral part of the United Nations, has no judicial powers to enforce the LPR system, and so responsibility for the implementation of the LPRs is delegated to the competent civil aviation authority (CAA) for each member state. To assist member states in their implementation, recommendations are published by ICAO in the second edition of the *Manual on the Implementation of ICAO Language Proficiency Requirements* (International Civil Aviation Organization 2010). This manual includes references to many commonly known areas of language assessment, including test development, test operation and administration.

The expected knowledge and skills of diverse stakeholders involved

The ICAO manual includes recommendations for the skills and knowledge required for certain stakeholders in the LPRs system, and a list of

¹ Language used in radio communication between pilots and ATCOs can be separated into 1) ‘Standard ICAO phraseology’ – *a specialised code of restricted sub-language for use in routine situations ensuring efficient and safe communications* – and 2) ‘Plain language’ which is defined as the *spontaneous, creative and non-coded use of a given natural language, constrained by the functions and topics (aviation and non-aviation) that are required by aeronautical radiotelephony communication* (International Civil Aviation Organization 2010; italics in original).

recommended *aviation language qualifications* is provided. The list includes categories of stakeholder as well as three levels of required criteria rated from *best* to *minimum*. At first sight, therefore, one may conclude that clear guidance and appropriate information about the knowledge and skills of language assessment required for each stakeholder group is offered, and that this information is comprehensive, appropriate, and fit for purpose.

On closer inspection, however, many of the recommendations for skills and knowledge given in the manual show a heavy reliance on language training rather than language testing, and provide only a limited reference to the requirements of subject matter experts (SMEs). This lack of any clear reference to the skills and knowledge required by those whose actual language proficiency is being assessed – the pilots and controllers – is a curious omission considering the importance mentioned earlier of TLU and indigenous criteria. One category of stakeholder that is not even mentioned is that of national authorities. Given their responsibility for administration of language proficiency test results in relation to pilot and controller licensing, as well as providing oversight and approval for language proficiency tests under their jurisdiction, this omission of a key stakeholder group is even more curious.

Recent issues raised by key stakeholder groups concerning the effectiveness of test instruments and their link with assessment literacy

Concern about elements of the LPR system has been evident since its implementation. Early fears concerned the deficiencies inherent in the ICAO Rating Scale, developed as a global tool providing criteria and levels for assessment of pilots' and controllers' language proficiency. It requires assessment of six language criteria (*Pronunciation, Fluency, Vocabulary, Structure, Comprehension* and *Interactions*) across six levels (1–6) with a minimum of (Operational) Level 4 in *all* six language criteria required to receive a pass result indicating achievement of a minimum level of language proficiency.

No empirical evidence was made available to substantiate the development or rationale of the scale and no one specific test was ever provided as a standardised instrument for use with the scale (Kim 2013, Knoch 2009). Such shortcomings and the subsequent inability of test service providers to supply evidence in support of their test instruments suggest inherent limitations within the system with clear implications for aviation safety (Alderson 2009, 2010). Another key issue has been the ambiguity over what exactly the assessment construct was. References to any sort of model construct suggested that it had either been conceptualised through a list of five holistic descriptors published in the ICAO documentation (Kim 2013), or embodied in the elements of the rating scale (Farris 2016). Indeed, Kim (2018) suggests

that certain skills and competences to be assessed in the rating scale, typically achieved via a traditional method of language assessment, may not actually be those skills required for real-life communication in aviation. Having a clear definition of the construct is recognised as one of the key initial stages in any test development process (Alderson, Clapham and Wall 1995, Fulcher 2010, Weir 2005). Interpretations as to what constitutes valid and appropriate test instruments have thus varied widely around the world, and the rating scale is often used as the *de facto* rationale to justify test validity and quality, albeit interpreted in widely differing ways (Bullock and Westbrook 2021, Knoch 2014).

More recent concerns have drawn on research into communicative competence, and highlighted an over-reliance on the system of language proficiency in aviation communication as an isolated measurable element in what has been described as a complex multi-disciplinary, and perhaps not always measurable, communication process (Bullock and Kay 2021, Monteiro and Bullock 2020). Doubts have thus been expressed that many current tests do not assess day-to-day communication needs and are thus of limited value (Elder, McNamara, Kim, Pill and Sato 2017, Kim 2013). Whilst this overview remains rather discouraging, the evolution of the LPR system has not been without attempts to tackle these issues. ICAO, in conjunction with partner organisations, has developed or supported a number of projects. A brief overview of some of these projects is given below.

Rated Speech Sample Training Aid (RSSTA):

Initially developed around 2010, the RSSTA project was devised to assist in rater/assessor training using a range of rated samples of speech taken directly from language proficiency exams. This training aid, although acknowledged as useful, is thought to be in need of updating due to the relatively short speech samples and doubts over the validity of certain test tasks used to produce the samples.

Aviation English Language Testing System (AELTS):

In 2011, ICAO introduced a system whereby providers of language testing services under the LPR system could have their test systems evaluated and, where compliant with standards of good testing practice, endorsed as such by ICAO. The project was criticised, however, for being too expensive – the application fee was \$7,000 – with little help or feedback offered to unsuccessful applicants. There was also a lack of transparency about the evaluation criteria and about who carried out the evaluation.

ICAO LPR Task Force (LPRI TF):

In 2005, the ICAO European and North Atlantic office, based in Paris, initiated an LPR Task Force (LPRI TF) and invited stakeholders to

regularly meet and discuss issues arising with the system. This group has met twice a year since, and includes CAA and Air Navigation Service Provider (ANSP)² representatives from member states, as well as representatives from professional associations. The task force has enabled the establishment of working groups to look at specific LPR issues, including the need for guidance material to assist stakeholders in their various roles.

Test Design Guidelines (TDGs):

In 2019, ICAO and the International Civil Aviation English Association (ICAEA) cooperated to deliver a series of worldwide workshops. These were designed to assist national civil aviation authorities and test service providers to better understand what design elements are considered essential (and why) when creating test instruments under the LPR system. This is particularly important for authorities as they have the responsibility for oversight and approval of tests under their jurisdiction.

Such projects clearly have a remedial function in addressing some of the concerns raised earlier and suggest that a certain level of awareness of the issues, hitherto absent, is now present within the system. However, it is still likely that for the LPR system to function effectively stakeholders need to be much more aware, not just of the skills and knowledge required for their own roles, but of the interdependence and shared understanding of all roles within the system (Bullock and Kay 2021).

Research into stakeholder groups' skills and knowledge

In early 2022 I carried out a small-scale research project in the form of a survey that addressed key stakeholder groups directly. This was undertaken in order to compile a list of the likely skills and knowledge required for each group, and to assess and evaluate the perceived importance attributed to each criterion; this could provide a platform for further in-depth research.

The project was the result of feedback and discussions that followed a webinar, organised by ICAO and delivered via Zoom, that I delivered with a colleague from ICAEA in October 2021. The webinar's objective was to summarise the role of language proficiency assessment in the process of aircrew (pilots) and ATCO licensing. Licensing was the principal focus of the webinar, as the language proficiency test result, and its validity, must be included as an endorsement in a pilot's or a controller's licence. Approximately 100 respondents from various locations and in various roles around the world attended the webinar, which focused on presenting some

² ANSPs are the organisations responsible for management and operation of a state's air traffic control services.

of the current issues within the LPR system as well as the connections and commonalities between stakeholder groups in the licensing process.

The survey

I carried out the survey shortly after the webinar in order to further explore some of the themes discussed during the event. The survey, conducted in the form of a questionnaire, was exploratory and interpretive in nature. It initially aimed to see what observations could be made about the perceived skills and knowledge required by each stakeholder group and the level of importance each group attaches to their roles under the LPRs. Secondly, I wanted to see if there were any possible connections between the levels of the skills and knowledge essential for each stakeholder group and some of the issues inherent within the LPR system. The survey was delivered online using the Qualtrics program.

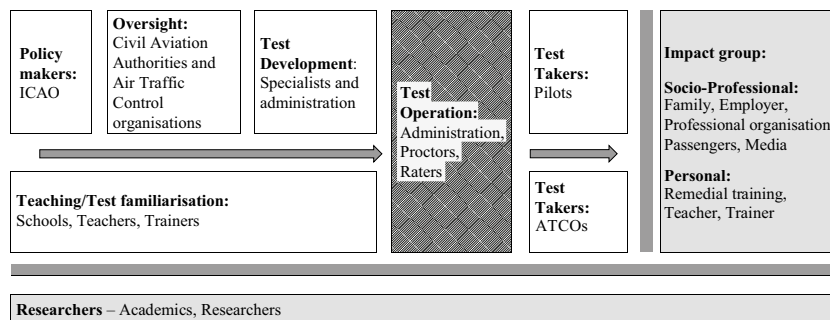
Stakeholder groups

The stakeholder categories in the questionnaire were defined by drawing on research that had previously attempted to explain specific roles within the framework of the LPRs (Douglas 2004, Knoch 2014) and was supported by an analysis of ICAO documentation (International Civil Aviation Organization 2009, 2010). The different stakeholder groups can be seen in Figure 1 below, which shows how they can be located within the process of the assessment system. The arrows show the sequencing of stakeholder involvement from policy making through oversight to actual test operation and the two main groups of test-takers. The *Impact group* in the final right-hand column was separated from the other groups as they are considered as *post-test* or *consequential* involvement. This simply means that these are groups *indirectly* involved in, or affected by, the system. Similarly, the group *Researchers* was separated from the main groups as they perform an indirect *external* and *overview* role, by observing and reporting on the system.

Whilst Figure 1 begins to provide clearer insights into the roles of the active stakeholder groups, it was also clear that many stakeholders have slightly varying roles (e.g. a pilot may be active in a private or a commercial capacity, a controller may work in a tower or an en-route centre). Furthermore, many stakeholders fulfil more than one role – a pilot, for example, could also be a rater, a controller acting as a test developer, and so on.

Because of such variation, I referred to information from the ICAEA membership, which comprises over 100 full-time members and more than 1,000 on an active mailing list. The information shows each member's primary (and secondary where available) role and the sector in which

Figure 1 Superordinate stakeholder groups in the LPR system process



they work. A final list was then compiled which included both primary and possible secondary roles, as shown in Table 1.

Table 1 Key stakeholder roles in the ICAO LPRs

1. Policy administrator (ICAO/EASA³)
2. CAA focal point/administrator⁴
3. ANSP focal point/administrator
4. Support association manager/administrator
5. Test developer/designer
6. Test centre administrator/proctor
7. Test examiner/interlocutor/rater
8. Air traffic controller
9. Pilot
10. Language teacher
11. Language school manager
12. Language teacher trainer
13. Researcher/academic
14. Other (please specify)

Stakeholder skills and knowledge related to LAL

Once the stakeholder groups had been identified, it was then necessary to compile a list of skills and knowledge related to the obligations for each group according to the LPR system. For this I drew on a range of ideas from both LAL theory and specific purpose language testing research (ALTE 2018, Bachman 1990, Douglas 2000, Monteiro and Bullock 2020). The ideas were formulated into a draft list which was then discussed and finalised with my

³ European Union Aviation Safety Agency

⁴ A *focal point* is the term used to refer to the person(s) designated as responsible for a particular task within the organisation.

ICAEA colleagues. The final list comprised 30 individual elements coded into five groups (see Table 2). Certain linguistic terms in the list were paraphrased to be as accessible as possible to a lay audience less aware of more traditional linguistic terms, e.g. Item 2.4 (*cut scores*) was glossed as ‘Rating scales and

Table 2 Skills and knowledge related to operational roles in the ICAO LPRs

1. Elements in language pedagogy
1.1 Language learning
1.2 Language proficiency
1.3 Communication skills
1.4 Intercultural awareness
2. Principles and concepts for testing (general)
2.1 History and theory of language testing
2.2 Validity and reliability
2.3 Authenticity of testing specific language
2.4 Rating scales and pass marks (cut scores)
2.5 Statistics
3. Features in test development and design (aviation)
3.1 Test development planning
3.2 Knowing what <i>exactly</i> is to be tested (construct)
3.3 Test delivery options (class/on-job/online)
3.4 Inclusion of both language <i>and</i> operational experts
3.5 Real-world communication (including radiotelephony)
3.6 Technical aeronautical knowledge
3.7 Separate tests for individual test-takers – pilots/ATCOs
3.8 Range of task types
3.9 Testing individual skills of listening and speaking
4. Aspects of test operation
4.1 Test operation and administration (general)
4.2 Issues with technology and equipment
4.3 Local testing infrastructure
4.4 Interlocutor/examiner behaviour
4.5 Variability between raters/assessors
5. Socio-professional factors
5.1 Testing language proficiency in aviation (ICAO Standards & Recommended Practices – SARP ⁵)
5.2 Inter-stakeholder awareness and communication
5.3 Inappropriate test use, cheating, misuse of scores
5.4 Test quality and usefulness
5.5 Social value and impact of results
5.6 Relationship with press/media
5.7 Communication with general public

5 The SARP^s is a safety management directive intended to assist states in managing aviation safety risks, in coordination with their service providers. It aims to support the continued evolution of a proactive strategy to improve safety performance and is based on the implementation of a specific State Safety Programme (SSP).

pass marks' and Item 3.2 (*construct*) was glossed as 'knowing what *exactly* is to be tested'.

The skills in the questionnaire could each then be judged on a level of importance by the respondents using Likert-type items with responses ranging from 1 (no skills/knowledge needed) to 5 (extensive skills/knowledge required). After completion of the survey, a total could be calculated for each skill across the stakeholder group, which was then divided to give a mean score per group, per skill, and then the responses would be converted to percentages. The higher the percentage, the more important an element was deemed to be. An example is given in Table 3.

Table 3 Example of level of importance given to each skills/knowledge criterion

Stakeholder Group	Skill	(From 1 to 5 based on Likert scale level)	(136/44)	(Percentage – x 20)
N = 44	History and theory of language testing	Total score = 136	Mean = 3.10	0.64

Once the questionnaire was complete, invitations to participate were sent out worldwide via LinkedIn and the ICAEA LinkedIn forum. The invitation was also sent via direct email to over 200 colleagues, with a request to forward to any internal or organisational mail groups where appropriate. A deadline of six weeks from the invitation being sent was set, in order to allow for timely completion of the questionnaire and retrieval of the data.

Questionnaire results

Respondents' roles

To begin examination of the data from the questionnaire, simple statistical analysis using Microsoft Excel spreadsheets was used, firstly to examine the roles of the respondents. For this analysis, only the 179 respondents who entered a primary role were considered. Of these, 45 noted a *primary role only*, whereas 134 entered both a *primary and a secondary role*. Seventy-eight percent ($n = 140$) of respondents who entered either a primary role only, or both a primary and secondary role, belonged to only five categories (see Tables 4a and 4b).

With such a large number of respondents belonging to just these five groups, and the potentially differing approaches that *operational* stakeholders and *language-related* stakeholders may have to their roles, it was decided to focus the subsequent analysis on the responses of these five principal groups

only. Furthermore, with the key importance attributed to experience in both *primary* (e.g. pilot) and *secondary* (e.g. rater) roles as mentioned earlier, the data from those respondents entering both *primary* **and** *secondary* roles will be considered as two individual responses. Thus, the total responses counted for the data analysis in the five groups given was 232 (see Table 4b).

Table 4a Respondents' roles in the five principal categories where primary roles were given

Responses	<i>n</i>	Percentage of all respondents by role	Percentage of all respondents
Oral examiner/assessor/rater	42	23%	78%
Air traffic controller	28	16%	
Language teacher	28	16%	
CAA focal point	23	13%	
Pilot	19	11%	
Total	140		

Table 4b Respondents' roles in the five principal categories where both primary *only* and primary plus secondary roles were given

Responses	<i>n</i>	Percentage of all respondents by role	Percentage of all respondents
Oral examiner/assessor/rater	83	27%	74%
Language teacher	51	16%	
Air traffic controller	36	12%	
Pilot	35	11%	
CAA focal point	27	9%	
Total	232		

Geographical location and first language

The majority of respondents came from the ICAO European and North Atlantic (EUR/NAT) region at 57% ($n = 102$), followed by Asia Pacific, 19% ($n = 34$) and South America, at 13% ($n = 24$). Twenty-six percent of respondents noted English as their first and main language (L1), although this does not appear to be related to the region in which they live. A variety of non-English L1s was noted.

Experience of the ICAO LPRs and exam environments

Eighty-one percent of respondents ($n = 126$) reported themselves as being moderately to very experienced (6+ years in their role) in their primary role and 72% of respondents ($n = 92$) claimed to have a similar amount of experience in their secondary role. In terms of the experience respondents

had of exams, 70% ($n = 123$) confirmed having been involved in conducting and/or administering LPR tests or exams while 55% of respondents ($n = 96$) stated having been tested, either as a pilot or an ATCO. Of those tested, a large majority, 78% ($n = 76$), stated having been tested face-to-face in a class or exam room, as opposed to in an operational setting or a remotely operated test. Seventy-three percent of respondents ($n = 125$) confirmed having attended formal training for testing and assessment, although 49% of those were in only three categories (rater training, oral examining, and general principles of testing and assessment).

Perceived importance given by respondents to all criteria

In order to identify and address the level of importance that each group attributed to all skills for their particular group, the 30 skills/knowledge elements were listed across the five stakeholder groups with the importance given in percentages. The data are presented in alphabetical order across each group in Table 5. The five groups are identified in the table as:

1. ELE (TE) – those respondents involved in teaching roles
2. ELE (TS) – those respondents involved in testing roles
3. SME (P) – pilots (all categories)
4. SME (A) – air traffic controller (all categories)
5. CAA – those involved in a national CAA role

The table includes the divergence (DIV) between the highest and lowest percentages and the mean percentage across all five groups for each skills/knowledge element. It also includes the divergence between the highest and lowest percentages and the mean percentage ratings across all skills/knowledge elements for each group.

The data in Table 5 shows a moderate to high level of importance attributed to all criteria, and this corresponds to observations made regarding self-reporting surveys in that those who do respond often deem all criteria of relatively high importance (Coombe, Vafadar and Mohebbi 2020, Malone 2013). Thus, there was no significant divergence observed between groups for each criterion, and minimal divergence for each group across all criteria. Across groups, *Rating scales and pass marks (cut scores)* showed the largest divergence of 0.27 (0.94 for testers and 0.67 for pilots). The teachers group showed the highest divergence across all criteria of 0.51 (0.93 for *Testing language proficiency in aviation (ICAO SARPS)* and *Language proficiency* with 0.41 given to *Relationship with press/media*). Some criteria were notably less important across all groups, such as *Statistics* (0.56), *Communication with the general public* (0.51) and *Relationship with press/media* (0.48).

Table 5 Importance given for all elements by each stakeholder group

Skill/Knowledge	Stakeholder groups						
	ELE (TE)	ELE (TS)	SME (P)	SME (A)	CAA	DIV	MEAN
	<i>n</i> = 83	<i>n</i> = 51	<i>n</i> = 35	<i>n</i> = 36	<i>n</i> = 27		
Communication skills	0.90	0.95	0.87	0.86	0.81	0.14	0.88
Language proficiency	0.93	0.97	0.83	0.86	0.81	0.16	0.88
Testing language proficiency in aviation (ICAO SARPS)	0.93	0.95	0.78	0.84	0.73	0.23	0.85
Knowing what <i>exactly</i> is to be tested (construct)	0.89	0.93	0.76	0.81	0.82	0.18	0.84
Real-world communication (including radiotelephony)	0.87	0.87	0.88	0.89	0.69	0.20	0.84
Interlocutor/examiner behaviour	0.88	0.95	0.78	0.78	0.71	0.24	0.82
Test quality and usefulness	0.85	0.89	0.79	0.82	0.72	0.17	0.81
Rating scales and pass marks (cut scores)	0.88	0.94	0.67	0.76	0.74	0.27	0.80
Testing individual skills of listening and speaking	0.82	0.90	0.83	0.75	0.69	0.21	0.80
Inclusion of both language <i>and</i> operational experts	0.79	0.86	0.79	0.78	0.69	0.17	0.78
Language learning	0.88	0.85	0.67	0.80	0.67	0.22	0.77
Authenticity of testing specific language	0.81	0.89	0.72	0.70	0.69	0.20	0.76
Validity and reliability	0.81	0.89	0.70	0.72	0.71	0.19	0.76
Inappropriate test use, cheating, misuse of scores	0.75	0.83	0.74	0.63	0.81	0.20	0.75
Intercultural awareness	0.82	0.84	0.70	0.72	0.67	0.17	0.75
Technical aeronautical knowledge	0.77	0.82	0.72	0.78	0.64	0.18	0.75
Variability between raters/assessors	0.80	0.84	0.71	0.71	0.71	0.13	0.75
Range of task types	0.79	0.84	0.66	0.68	0.69	0.18	0.73
Separate tests for individual test-takers – pilots/ATCOs	0.77	0.75	0.63	0.75	0.77	0.14	0.73
Social value and impact of results	0.74	0.74	0.70	0.70	0.74	0.06	0.73
Inter-stakeholder awareness and communication	0.76	0.77	0.64	0.73	0.68	0.13	0.71
Test delivery options (class/on-job/online)	0.66	0.74	0.68	0.74	0.70	0.08	0.70
Test operation and administration (general)	0.73	0.73	0.58	0.67	0.68	0.15	0.68
Issues with technology and equipment	0.66	0.73	0.60	0.68	0.64	0.13	0.66
Local testing infrastructure	0.62	0.72	0.56	0.73	0.65	0.17	0.66

Test development planning	0.66	0.67	0.48	0.65	0.68	0.21	0.63
History and theory of language testing	0.62	0.64	0.52	0.54	0.63	0.13	0.59
Statistics	0.54	0.58	0.43	0.58	0.66	0.23	0.56
Communication with general public	0.48	0.50	0.53	0.56	0.49	0.09	0.51
Relationship with press/media	0.41	0.49	0.45	0.56	0.48	0.15	0.48
Highest	0.93	0.97	0.88	0.89	0.82		
Lowest	0.41	0.49	0.43	0.54	0.48		
Divergence	0.51	0.48	0.45	0.36	0.34		
Mean	0.76	0.80	0.68	0.73	0.69		

Similarities and differences between stakeholder sub-groups

Because of the particular interest in specific purpose language in the roles played by both operational (SMEs – in aviation, the pilots and ATCOs) and language experts (ELEs – the teachers and testers), there is an interest in observing the response data of these two specific groups. Both SME groups agreed on the same eight out of the 10 most important skills and knowledge. Indeed, both groups of SMEs rated *Real-world communication (including radiotelephony)*, *Communication skills*, and *Language proficiency* as the three highest. The ELEs, likewise, ranked the same eight skills in their top 10, and like the SMEs, each sub-group of ELEs rated the same top three in terms of importance: *Language proficiency*, *Testing language proficiency in aviation (ICAO SARPS)* and *Communication skills*.

One omission of note in the top 10 most important criteria for the testers group was *Real-world communication (including radiotelephony)*. Given the level of importance attributed to authenticity and communication from the real world in specific purpose language testing (Bachman and Palmer 1996, Douglas 2000, Moder 2013, Yan 2009), this observation may seem surprising. Whether this particular criterion was misinterpreted is uncertain, although the addition of *radiotelephony* may have caused some confusion. ICAO documentation does, in fact, suggest that ‘The linguistic analysis of phraseology ... belongs solely in the hands of qualified operational personnel’ (International Civil Aviation Organization 2010:3.5). At the same time, however, the same documentation is quite clear when ICAO stipulates that: ‘The holistic descriptors and Rating Scale were **developed to address the specific requirements of radiotelephony communications**’ (2010:6.9; author’s emphasis) and ‘It is important that flight crews and air traffic controllers be proficient in the use of plain language used **within the context of radiotelephony communications**’ (2010:6.11; author’s emphasis).

Thus, despite the inclusion of standard phraseology in any pilot/ATC radiotelephony communication, many stakeholders may have chosen to interpret the LPRs as strictly related to plain language alone. Further research is certainly recommended here to learn more about stakeholders’ understandings of these terms as well as perhaps providing clearer guidance on the constituent elements of radiotelephony communications in a test construct.

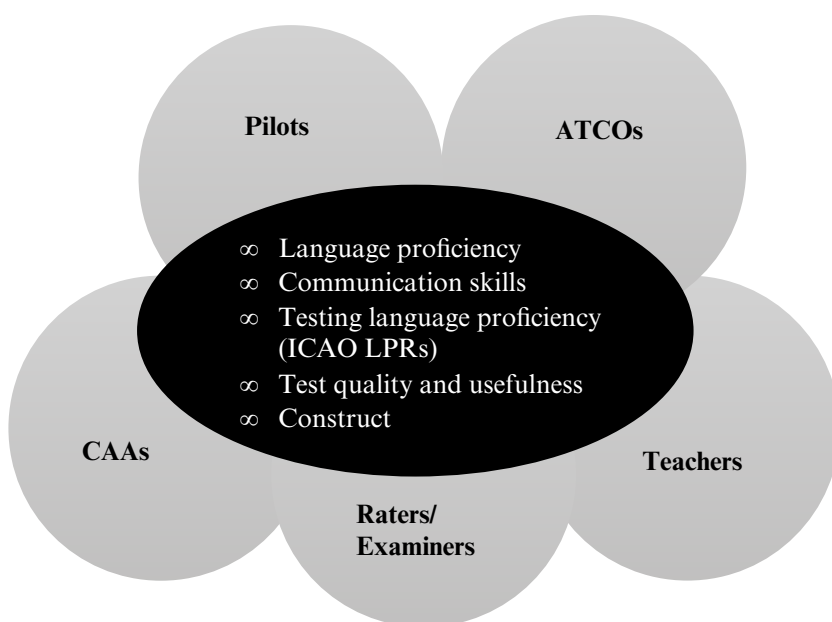
Discussion

The observations here provide an overview of the kinds of skills and knowledge that can be included to raise awareness of assessment literacy in aviation, although it is too early to formulate specific recommendations from such initial data. When comparing all five main stakeholder groups, the same

five skills and knowledge areas were included in the 10 highest ranked of each group. Figure 2 shows these five skills in the central circle. Three of the skills areas (*Language proficiency*, *Communication skills*, and the recommendations from ICAO on the *LPRs* and *SARPs*) can be seen as extremely reassuring. The LPR system is after all based on assessment and maintenance of an operational level of language proficiency which is part of the wider group of communication skills. The ICAO system and its SARPs are the means by which this is operationalised. The notion of *Communication skills* is also laudable by its inclusion in that it addresses a wider set of skills than simply language in isolation. The omission by some groups of *Real-world communication* knowledge is surprising, given the fact that communication in a test-taker's operational setting, as mentioned earlier, should be included if an assessment is to be made of those same skills. To what extent this can be attributed to individual interpretations of criteria by different respondents requires further investigation.

The final item included by all groups in their top 10 'most important' rankings, *Test quality and usefulness*, is also of interest as it may be linked to those stakeholders who raised doubts about the quality of tests being used to assess language proficiency in aviation. The concept of quality is, of course, very subjective. Quality to one stakeholder may relate to price, speed and

Figure 2 Skills and knowledge judged important across the main groups



efficiency, with lesser regard for adherence to good testing practice, whereas another may see the complete opposite. Further direct observations are therefore needed to supplement such reflections and reach more concrete conclusions.

The data show that there are several common judgements of the skills and knowledge required by key stakeholders. What is encouraging is that the stakeholder groups appeared to confirm a similar level of importance of many of those principal skills and knowledge, both in terms of what are more and what are less important. At this point it is maybe premature to make further conclusions, and further research is certainly recommended in order to determine why some areas were deemed more and some less important. Such research could also hopefully help evaluate the importance of the skills and knowledge against the requirements of each stakeholder role. The results of which could then be developed into a workable framework providing a clearer concept of interdependency for all stakeholders in this domain.

What is less evident from this initial research project is how any proposed increase in knowledge and skills awareness could better address emerging issues within the system of the LPRs. Whilst certain hypotheses are possible, such preliminary research is unable to sufficiently demonstrate clear connections between problematic areas and a lack of assessment literacy. Such observations do, however, provide a basis for scoping research to explore the direct association between improving the concept and practice of LAL in this domain and finding solutions to the issues highlighted.

Conclusion

This chapter aimed to highlight emerging issues in the domain of testing language proficiency in aeronautical communication worldwide, and set out to discuss observations from research involving over 170 key stakeholders involved in the LPR system. The research intended to see what commonalities and potential discrepancies existed between stakeholders on the perceived importance of skills and knowledge within the system. The resultant survey data showed that a certain level of agreement exists on the importance of those skills and knowledge required. This project has therefore provided a basis for more in-depth research into LAL across all stakeholder groups in this specific purpose domain.

Further investigation through qualitative research into stakeholders' conceptualisations of these terms in the survey would certainly be worthwhile in order to better judge the extent to which their perceptions correspond with and differ from each other's and why. Research questions could also be asked about perceptions and inherent reasons on the use of real-world communication in testing, and to better identify what constitutes test quality in the view of all stakeholders. Additional studies could aim to determine why

stakeholders perceive interaction with outside groups, indirectly involved in the testing system, with relatively low importance. Taken together the results could then be used to suggest how best to integrate required skills and knowledge into developing possible training programmes or guidance material that focus on the needs of all stakeholders.

Such development of knowledge and ideas could therefore provide an effective way of allowing the language proficiency system in aviation to evolve in an interdependent, efficient and effective way, through greater understanding between stakeholders. At the same time, such an evolution would continue to focus on aviation safety by a greater awareness of language proficiency in real-world communication and by the development and delivery of appropriate and quality-based assessment instruments for pilots and controllers worldwide.

References

- Alderson, C J, Clapham, C and Wall, D (1995) *Language Test Construction and Evaluation*, Cambridge: Cambridge University Press.
- Alderson, C J (2009) Air safety, language assessment policy, and policy implementation: The case of aviation English, *Annual Review of Applied Linguistics* 29, 168–187.
- Alderson, C J (2010) A survey of aviation English tests, *Language Testing* 27 (1), 51–72.
- Association of Language Testers in Europe (ALTE) (2018) *Guidelines for the development of languages for specific purposes tests – A supplement to the manual for language test development and examining*, available online: www.alte.org/resources/Documents/6093%20LSP%20Supplement%20-%20WEB.pdf
- Bachman, L F (1990) *Fundamental Considerations in Language Testing*, Oxford: Oxford University Press.
- Bachman, L F and Palmer, S (1996) *Language Testing in Practice*, Oxford: Oxford University Press.
- Baker, B (2016) Language assessment literacy as professional competence: The case of Canadian admissions decision makers, *Canadian Journal of Applied Linguistics* 19 (1), 63–83.
- Bullock, N (2015) Defining meaningful material for the teaching of English for aeronautical communications, in Borowska, A A and Enright, A (Ed) *Changing Perspectives on Aviation English Training*, Warsaw: Uniwersytet Warszawski, 35–45.
- Bullock, N and Kay, M (2021) *Pilot and ATCO licensing – The ICAO Language Proficiency Requirements*, ICAO Webinar presentation, 26 October 2021.
- Bullock, N and Westbrook, C (2021) Testing in ESP: Approaches and challenges in aviation and maritime English, in Lanteigne, B, Coombe, C and Brown, J D (Eds) *Challenges in Language Testing Around the World – Insights for Language Test Users*, Singapore: Springer, 67–77.
- 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 (1), 1–16.

- Douglas, D (2000) *Assessing Language for Specific Purposes*, Cambridge: Cambridge University Press.
- Douglas, D (2004) Assessing the language of international civil aviation: Issues of validity and impact, in IEEE Professional Communication Society (Ed) *International Professional Communication Conference*, Minneapolis: IEEE Professional Communication Society, 248–252.
- Dudley-Evans, T and St John, M (1998) *Developments in ESP*, Cambridge: Cambridge University Press.
- Elder, C, McNamara, T, Kim, H, Pill, J and Sato, T (2017) Interrogating the construct of communicative competence in language assessment contexts: What the non-language specialist can tell us, *Language & Communication* 57, 14–21.
- Farris, C (2016) Aviation language testing, in Estival, D, Farris, C and Molesworth, B (Eds) *Aviation English: A Lingua Franca for Pilots and Air Traffic Controllers*, New York: Routledge, 79–91.
- Fulcher, G (2010) *Practical Language Testing*, London: Hodder Education.
- Fulcher, G (2012) Assessment literacy for the language classroom, *Language Assessment Quarterly* 9 (2), 113–132.
- Harding, L and Kremmel, B (2016) Teacher assessment literacy and professional development, in Tsagari, D and Banerjee, J (Eds) *Handbook of Second Language Assessment*, Handbooks of Applied Linguistics Volume 12, Berlin: Mouton de Gruyter, 413–428.
- Inbar-Lourie, O (2008) Building literacy in language assessment competencies: A focus on language testing courses, *Language Testing* 25 (3), 385–402.
- Inbar-Lourie, O (2013) Guest editorial to the special issue on language assessment literacy, *Language Testing* 30 (3), 301–307.
- Ingham, K and Thighe, D (2006) Issues with developing a test in LSP: The International Certificate in Financial English, *Research Notes* 25, 5–9.
- International Civil Aviation Organization (ICAO) (2009) *Cir 323 Guidelines for aviation English training programmes*, Montreal: ICAO.
- International Civil Aviation Organization (ICAO) (2010) *Doc 9835 Manual on the implementation of ICAO language proficiency requirements* (Second edition), Montreal: ICAO.
- Jacoby, S and McNamara, T (1999) Locating competence, *English for Specific Purposes* 18, 213–241.
- Kim, H (2013) Exploring the construct of radiotelephony communication: A critique of the ICAO English testing policy from the perspective of Korean aviation experts, *Papers in Language Testing and Assessment* 2 (2), 103–110.
- Kim, H (2018) What constitutes professional communication in aviation: Is language proficiency enough for testing purposes?, *Language Testing* 35 (3), 403–426.
- Knoch, U (2009) *Collaborating with ESP stakeholders in rating scale validation: The case of the ICAO rating scale*, Spaan Fellow Working Papers 7, Michigan: English Language Institute, University of Michigan.
- Knoch, U (2014) Using subject specialists to validate an ESP rating scale: The case of the International Civil Aviation Organization (ICAO) rating scale, *English for Specific Purposes* 33, 77–86.
- Malone, M E (2013) The essentials of assessment literacy: Contrasts between testers and users, *Language Testing* 30 (3), 329–344.
- Moder, C L (2013) Aviation English, in Paltridge, B and Starfield, S (Eds) *The Handbook of English for Specific Purposes*, Chichester: John Wiley & Sons, 227–242.

- Monteiro, A and Bullock, N (2020) A broader view of communicative competence for aeronautical communications: Implications for teaching and high-stakes testing, *ESpecialist* 41 (3), 2–29.
- Taylor, L (2009) Developing assessment literacy, *Annual Review of Applied Linguistics* 29, 21–36.
- Taylor, L (2013) Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections, *Language testing* 30 (3), 403–412.
- Weir, C J (2005) *Language Testing and Validation: An Evidence Based Approach*, Basingstoke: Palgrave Macmillan.
- Yan, R (2009) *Assessing English Language Proficiency in International Aviation*, Saarbrücken: VDM.

9

Shaping a language testing curriculum: Insights from an oral history of a Master's programme

*John Pill, John Bandman, Raffaella Bottini,
Tineke Brunfaut, Natalya Davidson, Geisa
Dávila Pérez, Luke Harding, Yejin Jung, Santi
B. Lestari, Camilo Ramos Gálvez, Olena Rossi
Members of the Language Testing Research Group,
Lancaster University, UK*

For J. Charles Alderson and Dianne Wall

This chapter provides an account of how the curriculum of a part-time, online Master's programme in language testing, delivered by distance, was conceptualised and modified over almost two decades, and the extent to which elements of the programme impacted on alumni's professional practice in language assessment. We draw on data collected as part of a wider oral history project, and specifically reflect on:

- The range of considerations guiding the original development of the programme's structure and pedagogical approach
- Factors influencing subsequent changes to the programme's content and delivery
- The programme's role in developing the knowledge and skills of emerging language testing specialists
- The impact of learning on students' subsequent professional practice

Background

A specialised Master's programme in language testing, the *MA in Language Testing (Distance)*, has been offered by Lancaster University for more than 15 years. Its format – part-time over two years and delivered online so that students can continue to work while they

study – was intended to allow access to professional development for applicants around the world who require knowledge and skills in the field of language testing and assessment, either to perform their current roles effectively or to progress to new professional and academic posts. So far, over 200 students have taken the programme for a range of purposes. For example, they teach a variety of languages to learners in primary, secondary, tertiary, and adult education in public and private institutions; participate in test development for their own school or university, or for national administration; develop and administer tests for pilots, the military, diplomats, oil platform or call-centre staff; and are employed by ministries and cultural organisations, universities and language schools, test providers and educational publishers. Graduates have often found the Lancaster Master’s qualification to be their entry into a professional community of language assessment specialists.

During its lifetime, the programme of study has undergone changes to its scope and content. However, course developers and teachers have not recorded these modifications in detail and, consequently, opportunities may have been missed to learn from reflecting more deliberately on the origins of the programme, subsequent changes, and their causes and effects. To remedy this, 11 members of the Language Testing Research Group – comprising academic staff and doctoral students at Lancaster University – initiated an oral history project in 2021. The project sought interviews with former and present course developers, administrators and support staff, teachers, and former students, and complemented these with a review of official documentation where available.

The project responds to a call in the field of applied linguistics for a greater focus on historical research, which Smith (2016) has termed *applied linguistics historiography*. According to Smith, existing histories of applied linguistics have tended to focus on the ‘development of theories in the abstract rather than paying much attention to practice’ (2016:79). Smith suggests that this is best addressed by histories which attempt ‘to situate ideas in particular contexts of practice (asking “what gave rise to them?”) and to ascertain their impact on practice, for example, on policies, on learning materials, or (most difficult, admittedly, to establish) on teaching and learning activities in particular settings’ (2016:79; see also Taylor and Green 2020).

We anticipate that a historical account of the MA in Language Testing – and particularly an oral history approach – will be of wider interest to the field of language testing and assessment. Previous research on language assessment courses in university settings has typically used surveys or interviews to explore issues related to curriculum decisions on standalone courses (e.g., Bailey and Brown 1996, Brown and Bailey 2008, Fulcher 2012, Inbar-Lourie 2008, Jeong 2013). While such studies have been crucial in building a broad understanding of good practice in course design, an

exploration of the MA in Language Testing provides unique insights into the nature of a *programme*-level curriculum for Master's education. Furthermore, an oral history perspective discloses not only what is taught on the programme but also how it is taught, why certain curriculum decisions were made, how the programme has evolved, and what aspects of learning on the programme have had resonance in the professional practice of former students. Fundamentally, oral history gives insight into the thinking of key informants at different points in time and allows us to consider the nature of changes over time. It therefore provides a methodology for exploring the complex intersection of historical, social, and institutional factors likely to exert influence on course design.

Methodology

The oral history approach used in this study was developed following the recommendations of key literature (Peniston-Bird 2009, Ritchie 2014, Thompson and Bornat 2017). Ethical approval was given by the Faculty of Arts and Social Sciences Research Ethics Committee at Lancaster University. Semi-structured interviews were conducted over a period of two months. Initial planning for data collection involved defining a sampling frame to recruit participants who had been directly involved in the MA in Language Testing (Distance) (MALTD) at Lancaster University since its start in 2007, as either students or staff members working on programme design, organisation, and delivery. Table 1 summarises the number of participants and their role.

A total of 28 possible participants were contacted and 23 agreed to be interviewed (11 females and 12 males). All academic staff agreed to be named interviewees in the study: Charles Alderson, Dianne Wall, Jayanti Banerjee, Tineke Brunfaut, Luke Harding, Judit Kormos, and John Pill. All taught on the programme, with four also having been director of studies (Alderson, Brunfaut, Harding, Pill) and four having been involved in the initial programme development (Alderson, Banerjee, Kormos, Wall). Interviewees also included Elaine Heron, the administrative coordinator, and Stephanie (Stef) Strong, responsible for developing the virtual learning environment

Table 1 Number of participants and their role

Participant role	No. of participants contacted	No. of participants interviewed
Academic staff	7	7
Administrative staff	2	2
Students	19	14
Total	28	23

and technical support. The group of former students – anonymous interviewees¹ – was diverse in terms of nationality. They also reported being based in one of a range of countries at the time of their studies: Austria, Chile, China, Czech Republic, France, Germany, Greece, Norway, Spain, Switzerland, and the UK. Additionally, they represented different cohorts: interviews were conducted with alumni from eight of the 13 cohorts that have graduated from the programme at the time of writing this chapter; interviewees came from Cohorts 2 to 12.

Interviews were conducted online using videoconferencing software (Microsoft Teams) and were audio- or video-recorded depending on the consent given. Each interview lasted about 60 minutes and was conducted by two members of the research team. Three of this chapter's authors not only conducted interviews but were also interviewees. Interview questions were shared with participants in advance. Different sets of questions were designed based on interviewee role (academic staff, administrative staff, student; see the Appendix for a sample of questions). Initial transcripts were obtained using automatic speech recognition software (Microsoft Stream). Each text was then manually checked for accuracy by a research team member. The 23 transcripts were double coded by team members to identify themes and patterns, specifically focusing on:

- recollections of curriculum matters (programme structure, sequencing, content, pedagogical approach) and the influences on these
- recollections of the impact of learning on students' subsequent professional practice.

A 'reconstructive mode of analysis' (Quinn, Chater and Morrison 2020:504) within the oral history approach (Thompson and Bornat 2017) was followed to identify similarities and differences across transcripts, and to develop an understanding of the narrative around the development of, and changes to, the programme.

Developing the programme: The early years

The origins of the MALTD are in the late 1990s, when Charles Alderson and Dianne Wall – academics at Lancaster University – were asked to set up a training programme for a British Council Peacekeeping English Project, following the fall of the Berlin Wall in 1989 and the restructuring of countries in Central and Eastern Europe. The Peacekeeping Project aimed to help language teachers and testers in the military of several countries that aspired to join NATO, as a NATO requirement was that military personnel

¹ To maintain student anonymity, numbers are used to refer to interviewees, e.g., Student_1. A student's cohort is indicated in parentheses, e.g., (C2) for Cohort 2.

were sufficiently proficient in English to ‘inter-operate’ in NATO exercises (Green and Wall 2005). Alderson and Wall had conducted language testing training activities throughout the 1990s in the Baltic States, Hungary, and the Slovak Republic. They were therefore well placed to respond to an invitation to deliver a custom-made programme for the Peacekeeping Project.

Dianne Wall: It was people who were working for the British Council in different countries in Central and Eastern Europe who needed help with testing. And so, they wrote to us. We didn’t write to them. But we catered for their needs and desires and eventually the original MA was established, which lasted for three cohorts.²

The resulting course – a precursor to the MALTD – was like some other programmes running at the time in Lancaster’s Department of Linguistics and Modern English Language (as it was then known). Modules were delivered through residential visits to Lancaster and supervision was conducted by distance, primarily using email communication. This original ‘Peacekeeping MA’ demonstrated the need for Master’s-level qualifications in language testing.

Dianne Wall: It just brought it home to us that there was a potential audience out there. There were large numbers of people in Europe, especially at that time in “new Europe” ... who needed and wanted training ... And that helped us to believe that it was a worthwhile project and that a few years later it would be worthwhile to expand the programme from the military context where it started to ... all contexts.

Preparations for what would become the MALTD therefore began in the early 2000s. An important initial consideration in developing the curriculum was the selection and sequencing of modules. The MALTD has always had the same structure, i.e., five taught modules with coursework, followed by a dissertation. Table 2 gives the original module names and briefly describes their content. The first three modules are taken in the first year, and the remaining modules and dissertation are completed in the second.

According to the original MALTD course designers – Alderson, Banerjee and Wall – the selection of modules was guided by pedagogical, institutional,

2 All direct quotations from participants in this chapter are modified from the raw transcripts – which naturally contain hesitations, fillers and repetitions – to assist with readability.

Table 2 Overview of the programme in the early years

Sequence	Module name	Brief description of module content
1	<i>Language Test Construction and Evaluation</i>	Evaluating and revising test items; test specifications; basic item and test statistics; testing the four skills.
2	<i>Background to Applied Linguistics for Language Testing</i>	Overview of applied linguistics; theories of language acquisition and use; theorising key language constructs.
3	<i>Issues in Language Testing</i>	Traditional and 'new' views on validity; rotating set of topics in language testing.
4	<i>Research Issues in Applied Linguistics</i>	Overviews of common research methods; qualitative data analysis; designing a research project; research ethics.
5	<i>Statistical Analysis for Language Testing</i>	Reviewing basic statistics; hypotheses and significance testing; inferential statistics; Rasch measurement.
6	<i>Dissertation</i>	One-on-one supervision; preparing a 15,000-word research dissertation.

and pragmatic factors. Two modules were already delivered on Lancaster's existing (on-campus) MA in Teaching English to Speakers of Other Languages (MATESOL) programme:

Jayanti Banerjee: Within the in-person MA [MATESOL] we had two language testing-focused courses: ... *Language Test Construction and Evaluation* – which was a whistle-stop tour through and an introduction to assessment – and then there was *Issues in Language Testing* – which allowed the MA students to dig into specific issues ... We wanted very much to make sure we covered those two sort of approaches into language assessment within the MA [MALTD].

The decision to start with *Language Test Construction and Evaluation* – a practical introduction to test design (based on Alderson, Clapham and Wall 1995) – was deliberate and a lesson learned from the Peacekeeping Project MA. Practical tasks such as critiquing poorly written items were viewed as a useful entry point to more demanding theoretical material:

Dianne Wall: An easy way in was by looking at things, the smaller scale, and that would be more relevant because all of these people were item writers, they were all test constructors. So, it's going from the familiar to the unfamiliar, if you will.

The inclusion of *Background to Applied Linguistics for Language Testing* as the second module appears to have been influenced by gaps in participants' prior learning.

Dianne Wall: *Background to Applied Linguistics* came directly from the original MA with the Peacekeeping network, because we saw that ... the people who were doing that, who had been invited or commanded to do testing in those contexts, were always teachers ... but they had studied languages in university, they hadn't necessarily studied linguistics or what is language or what is language learning ... So, it became apparent as those cohorts went on that we needed more applied linguistics ... and it needed to be in there early.

That the programme was located within a department of linguistics, rather than, say, a department of education, also influenced the prominence given to this module. The original course designers needed to demonstrate to departmental committees that the degree was congruent with other departmental offerings.

Jayanti Banerjee: We were a linguistics department, and this course was a Language Testing MA and there were questions around "Well, where's the linguistics in it?" ... Which is fair. Really we should know a good amount about language if we're going to assess it ... The MA gives an opportunity perhaps to build out people's understanding.

Issues in Language Testing was placed third, while *Research Issues in Applied Linguistics* was included as the fourth module, to prepare students for their dissertation and other research in their professional contexts. The inclusion of *Statistical Analysis for Language Testing* as the final coursework module is noteworthy; while basic statistics such as facility values and discrimination indices were introduced in the first module, the course designers believed that graduates of an MA degree in language testing required more.

Jayanti Banerjee: If you're going to be any kind of decent language assessment professional, be it a researcher or a creator of tests, you have to be able to do statistics ... And in fact, if there's anything that was a gap in our language assessment provision up to that point, with the in-person [MATESOL] students, [it] was that they got a little bit of statistics, classical test statistics, but nothing particularly

high powered. And we felt that they needed a better grounding.

The need for a strong statistical foundation might also have been influenced by the annual Language Testing at Lancaster Summer School, started in 2001 and run by Rita Green³, which included substantial statistical content. The design for the initial MALTD can therefore be understood as a mix of modules: the pre-existing foundational kind, those designed to reflect the identity of the administering department, and the rest aiming to address the anticipated needs of the first cohorts of students.

An innovation agreed early on was that, rather than following the residential model of the Peacekeeping Project MA, the new MALTD would be run entirely in distance mode. This was an unusual decision in the early 2000s, particularly for a team with little experience of online teaching. However, it was felt there was no choice.

Dianne Wall: It had to be done because [the] audience out there ... were working people. And when the [Peacekeeping Project] funding stopped they could no longer come ... To go online would mean that we'd have access to all those people out there and they didn't have to leave or sacrifice much apart from time in order to get a postgraduate degree.

The need to then translate teaching methods and pedagogical principles to an online environment provided the greatest hurdle at the design stage:

Dianne Wall: The *innovation* was the methodology. So, setting tasks for people, getting them to respond, getting them to reflect, getting them to do group work, getting them to present ... trying to work out how to get all of this process-oriented, task-based, inductive, Socratic method, reflection, all of this into the course itself, given the technology we had.

Alderson and Wall explain that this pedagogical approach on the MALTD, which prioritised tasks, group work, reflection, and discovery, had a direct lineage to the innovations of communicative language teaching that prevailed throughout the 1980s and 1990s at Lancaster University.

3 Green later helped redesign the MALTD statistics module, leading to her book *Statistical Analyses for Language Testers* (2013).

Dianne Wall: We started off in ... the Institute for English Language Education [IELE], which was really a cutting-edge place in its time. So, it was like the home of communicative language teaching and communicative training and the like. And Charles [Alderson] started up there ... and all of these things that I mentioned to you were part of the fabric of that institution. That was the way people taught. And I when I joined the institute, I must admit, Charles was a great mentor. So, I had been taught that way. I learned to teach that way. Jayanti [Banerjee] became part of that as well. The department had some of that ... The Department of Linguistics, which was like the mother department to the IELE, had people in it who taught that way as well. Mike Breen, Chris Candlin, Dick Allwright ... So there was a group of people for whom this *was* teaching.

Once again, the academic milieu in which the course designers taught and were trained proved to be important. The MALTD can thus be viewed as a site of innovation, not just in terms of online language testing instruction, but in the application of this communicatively oriented pedagogy to a Master's-level professional degree in an online environment.

The technical work fell largely to Stef Strong, an IT technician hired to assist with course design who spent several years working at a desk in Alderson's office. As Wall notes: 'A lot of stuff we imagined, and then we gave it to Stef Strong.' Having begun using an in-house virtual learning environment (VLE), the team soon realised that Moodle – an external, open-source VLE developed around a social constructivist pedagogy – would be the most appropriate way to deliver the programme. Moodle had its own influence on the shape of the programme:

Stef Strong: I would say [Moodle] focused it a lot. It's made it quite structured. And it's probably developed into what it is like, the weekly structure and it gives it its flow and the way we've structured it into the week by week.

This paved the way for a 'typical' structure with weekly units and a series of tasks. Students would gain access to tasks and readings each Monday morning, and post and respond on discussion forums set up for each task. Tasks would include reflection activities (e.g., read the article and comment on how you would apply the ideas in your own context), discovery learning activities (e.g., use this rating scale to judge a set of writing performances, then comment on the strengths and limitations of the scale), and group work activities (e.g., develop a set of materials for a speaking test in a context where

you have few material resources and limited time to conduct the assessment). The discussion forums were the ‘meeting place’ for students and teachers, and detailed feedback was provided for each task.

Dianne Wall: I remember spending hours giving feedback because I was aware that I wasn’t only giving feedback, but that the writing, the way it was phrased, the way it was organised, was like writing mini-essays every week. And that was good training to actually think about how you give feedback so it’s coherent and not trivial and it all makes sense.

When launched in January 2007, the MALTD programme had an initial cohort of 13 students. The developers made changes during the initial iterations of the programme to deal with less anticipated challenges, for example, making materials accessible in terms of their file format and complying with copyright regulations.

Change and evolution

Interviews with staff revealed several factors affecting the subsequent development of the programme: staff profiles and the programme structure; developments in the wider field of language testing; student profiles; and technological advances in the programme delivery.

Staff profiles and programme structure

Following the set-up period, the MALTD underwent some important changes early on. First, there were considerable movements in academic staff. Jayanti Banerjee left Lancaster at the end of 2007 to join the University of Michigan English Language Institute.⁴ Judit Kormos began working at Lancaster University in 2008, followed by Andrea Révész and Tineke Brunfaut. All contributed to different modules on the programme, and Brunfaut played a significant part by taking the director of studies role until 2016. Luke Harding joined at the end of 2010, just prior to the retirements of Dianne Wall and then Charles Alderson. By the end of 2012, the programme designed by Alderson, Wall, and Banerjee had effectively been handed over to this new team. They would be joined in 2017 by John Pill, the current director of studies.

⁴ Banerjee was later involved in the MALTD again, teaching individual modules and supervising some student dissertation projects on a visiting basis.

As these changes in personnel took place, there was a relative degree of stability at the broad programme level:

Judit Kormos: Overall, I think the kind of general structure as far as I see of the programme hasn't changed too much, so the kind of the bricks, the key elements are still there.

The interviews suggest that the team of teachers who 'inherited' the programme valued its structure as a useful model for training in language testing and assessment:

Luke Harding: It's quite structured ... we have this practical element in the first module, constructs in the second module, a kind of critical issue-based assessment literacy in the third module, and then we move into research and statistical literacy.

Similarly, the pedagogical approach adopted by Alderson, Wall, and Banerjee has also been valued by more recent module teachers, particularly the emphasis on group work and online interaction around a common task, which has since informed the development of other distance programmes in the (renamed) Department of Linguistics and English Language.

Judit Kormos: That was a really innovative feature of the pedagogy that meant from the very very beginning ... interaction among students was built in. And the idea was that the students learn as much from interacting with each other as they learn from interacting with the materials and us [teachers].

At the same time, there has been change within this general structure over the lifetime of the MALTD. Brunfaut notes: 'We've continued to work on the programme and let it evolve with us, with the field, with the students.' That the programme evolved with new academic staff was inevitable. New teachers brought with them different interests and so influenced the programme content, while departures and retirements meant that some topics became less prominent.

Judit Kormos: When Luke joined us then there was much more ... emphasis ... on English as a *lingua franca*, World Englishes, the comprehensibility aspects of language assessment ... Tineke's expertise in working in standard setting in European countries, exam projects, then [John] joined us with his expertise on testing in professional contexts and then of course as Charles and Dianne retired, these [areas of]

expertise were brought to the front and perhaps some of the expertise that was there on washback got lost in this sense, that was something we were known very well for.

However, such changes were often relatively minor and manifested at the unit or task level:

Tineke Brunfaut: We've introduced themes we are more expert in as well, and replaced the materials or updated some materials, but not in the sense that we, each time a new member of staff came on, we got rid of a module and started from scratch, definitely not. It was always within the frame that already existed that people have brought in their personal touch, their expertise.

Developments in the wider field of language testing

Factors beyond staff changes have impacted on the content of the modules. Most prominently, developments in the wider field led to ongoing re-evaluation of programme content. In particular, the *Issues in Language Testing* module has remained in flux:

Tineke Brunfaut: Topics changed, essentially, on an annual basis, because that module is specifically looking at “hot themes” in the field and what students find interesting or what are considered to be pressing areas.

Within that module, both Brunfaut and Harding point to how initial units on validity theory have been adapted over the past 10 years, moving on from traditional views to include both argument-based approaches and the sociocognitive perspective. In all modules, coursebooks and readings have been frequently updated. However, in some cases core readings are maintained for their pedagogical value. Curating reading lists while publication in the field grows quickly is one of the ongoing challenges for module teachers. Pill worries that ‘the temptation is to keep adding to the materials and what’s available and running the risk of overwhelming everybody with content which nobody processes properly.’

A significant recent change to the programme concerns the *Background to Applied Linguistics for Language Testing* module, which was revised and renamed *Exploring Language Constructs for Language Testing* in 2019.

John Pill: [The new module title is] saying to the students, “Look, you might not think that areas of second language acquisition like

pragmatics and interactional competence are important, but unless you know about those things, you're not going to be able to write tests that include those things." ... The re-titling of it was a clever way of making it explicit to students that this is core stuff in language testing. It's not a peripheral add-on about background to anything ... This is essential to constructs, which are, you know, fundamental to any test.

The re-orientation of the module name was concurrent with revision to its content, bringing this module into line with wider developments in applied linguistics research.

Tineke Brunfaut: I changed the grammar unit, for example, to reflect more language-in-use approaches to grammar assessment and grammar teaching. Or we replaced the entire vocabulary unit. [For speaking,] we brought in some new constructs like intelligibility, fluency, as they gained more importance in the field and the literature over time.

As illustrated in the quote from Pill above, it seems that a focus on construct has always been a feature of the programme. This, again, can be attributed to the context in which the programme is administered:

Luke Harding: So, I do feel that kind of intense focus on the construct is something that is, I would say, somewhat related to the fact that we're in this linguistics department and we're surrounded by amazing scholars doing work on discourse analysis and speech science and corpus linguistics and SLA [second language acquisition], and all of this stuff that we're exposed to feeds into what we are able to offer to our students.

Student profiles

Another driver of change was the shifting profiles of students who enrolled on the programme.

Judit Kormos: I think our initial student population was very diverse, from very different parts of the world with very different needs in terms of testing. So, we had one student for example [who] was interested in the testing of infrequently spoken languages ... to the very specific needs of the Austrian Matura [secondary school-leaving exam] team who enrolled in the MA language testing programme. There were times

when we had people ... from Cambridge Assessment and from the British Council et cetera. So, we had this very wide range of students, students working in the aviation field, military, so in military language testing.

The changing profile of students has affected the programme more indirectly than changes to the design of modules have – it prompts growth in the professional knowledge of classmates and the teachers themselves through sharing of contextualised experiences and practices. The reflections of students in weekly discussions have formed a vital and ever-changing part of module input.

Luke Harding: You learn a huge amount just by interacting with these students because really you're getting first-hand accounts of testing practice, so the challenges of it, theory meets practice ... The students bring an awful amount of enriching knowledge and insight that we take away as teachers in this programme and that is shared within the cohort as well. And I think whenever I've spoken to students in the past who've graduated, I think so many have said that they just learned enormously from their peers.

Technological advances in programme delivery

Technological affordances have also had an impact on pedagogical choices in the delivery and content of the programme. In the early years, options were more constrained by the VLE being used, and this has changed over time. New functionalities have included, for example, wikis, a digital library service that interfaces with Moodle (allowing direct click-through to texts online), and the integration of Microsoft Teams for videoconferencing. A particular ongoing challenge is the shift from a focus on written text – a feature of the programme for many years – to multimodal input.

John Pill: People these days, for example, [are] completely familiar with ... video delivery of material and interaction via video, and until recently that hasn't been a major part of the content, [of] the materials of the course at all, because that was very high bandwidth, expensive to download and was avoided. But these issues have changed over time, these restrictions have been removed and everyone now is making videos on their mobile phone and so on. And we need to keep up with that, as people delivering information, ideas in a modern format and in an accessible format for modern students.

To summarise, interviews with teachers on the programme show a curriculum that has developed and evolved in relation to change in four main spheres: staff profiles, developments in the wider field of language testing, student profiles, and technological advancement for programme delivery.

Learning on the MA and its impact beyond: The student perspective

Interviews with former students provided a perspective on the learning experience that complemented the views of academic staff. They also showed how learning was applied in students' professional practice after completing the programme. One notable observation was how the fundamental assumption that drove the introduction of the programme – that there were people involved in language testing who sought a more theory-based understanding of their practice – proved to be correct, for students in both earlier and later cohorts:

Student_11 (C2): The MALTD programme ... came at a very relevant and important time because I was quickly finding that my practical experience was far outweighing my theoretical understanding of the issues in my testing project.

Student_6 (C10): The theoretical aspects ... the foundational knowledge in language testing in the programme, it's something quite unique, right? ... To have a two-year focus just on testing is really really rare. Having that background knowledge of testing now at my stage in my career ... being able to apply those theoretical parts, the academic parts, now those are the parts that I'm finding most useful.

Our analysis of alumni recollections to identify narratives of learning reveals three recurring themes: increased awareness of principles of good practice; better understanding of statistics and research methods; and changes in personal beliefs, especially regarding ethics and responsibilities. These are exemplified in the quotes from former students in the following three sections.

Increased awareness of principles of good practice

Students gained awareness of key principles of language testing which, as a result, bolstered professional confidence.

Student_3 (C7): [The programme] makes you more professional ... it enables you to speak with relative confidence at least about matters of language testing ... It's very useful to have all those convincing arguments and having learned how to reasonably present these or facilitate for people to understand the reasoning behind that, the purpose of doing things that particular way.

'Being professional' in this extract is associated with the ability to draw on a principled set of arguments to communicate effectively and justify decisions to different audiences. Other students recollected that their learning experience during the programme helped them to (further) develop their identity as assessment specialists, (re)affirming their commitment to promote good practice in their workplace.

Student_1 (C8): It's sort of a bit of a joke really that I am the voice at the management meetings that says, "I know we can't do that", "That's not possible", "Well, what about validity?" ... "What about the construct of it?" ... I guess before I might have let things go a little bit more but ... [now] I am more meticulous where language testing in our field is concerned.

Student_4 (C12): We had a project where a group of people were marking some scripts and I said, "Well, I need a senior examiner there to oversee and to do quality assurance checks" ... I actually got my way in the end. I said, "No, listen to me ... we need this person to do this because of, you know, reliability."

In these cases, former students challenged established views and habits in their workplaces by presenting theoretically grounded arguments and positioning themselves as experts. Some students reported that this engagement with principles had tangible effects:

Student_7 (C4): Authenticity is a good example of something that I hadn't really thought about before, but then to have the chance to think about it and then think about okay how would I put that into practice in a test? That was really influential, I think, on my thinking in language testing and in terms of the work that I do now when we're thinking about, I don't know, potentially new products or how we might change and develop our products.

In addition, interviewees noted the benefits of learning how to conduct research:

Student_2 (C5): What kind of research people can do ... I think that was sort of the most, the biggest evolution, I guess, across the two years programme, to get an idea of what solid research is or can be, and what sort of things are that you could find out ... how do you design a solid research study, how do you write it up, how do you present it in a way [that is] engaging and accessible?

Students reported that learning about the research process helped to address areas of their professional practice that previous academic study had not prepared them for.

Student_10 (C10): My French studies degree at [University_Name] didn't include anything on that side and I dropped basically all the science subjects including maths ... So that was particularly challenging, but also particularly necessary to start looking at things more statistically, and then also thinking about how you find things out, really, how you research things, and whether it's the quantitative stuff or the qualitative stuff and how you bring them together.

Better understanding of statistics and research methods

The terms *statistics*, *statistic*, *statistical* and *statistically* occur 67 times across 13 of the 14 alumni transcripts. Learning about statistics was described as scary, challenging, and boring, but ultimately necessary.

Student_3 (C7): I was reading a journal ... it was heavy on the statistics, and yeah, I mean, it wasn't riveting ... it was sort of [a] boring article. But what I did was I started studying all these nice tables that were in the article ... I realised that I know what these things mean ... that was a moment of revelation, and then I realised that this is extremely useful, I never thought I would be able to look at anything [related to] statistics.

The modules *Research Issues in Applied Linguistics* and *Statistical Analysis for Language Testing* were generally described by students as particularly useful for their professional development, to understand and critically evaluate existing research – as in the example above – as well as to design their own studies, and to develop and evaluate language tests.

Student_6 (C10): Changing that focus to a test as something that's analysable statistically and qualitatively. That was an aha moment. Like this is something that you can actually take data from, and understand something about this instrument, the quality of the instrument, the applicability.

These examples illustrate how the students' approach to research methods and statistical knowledge often changed during the programme.

Changes in personal beliefs, especially regarding ethics and responsibilities

Personal beliefs and attitudes to language assessment developed and were often re-evaluated. For instance, interviewees often mentioned gaining a different perspective on the role of language assessment in their local contexts or a new approach towards problematic situations.

Student_9 (C5): The doubts about the ethics or the behaviour in my country was reinforced a little bit, and not only reinforced but I had the opportunity to try to look for the solutions. So, that was good.

Although increased awareness of ethical aspects of language assessment raised concerns in some contexts, it also provided an opportunity for students to reflect on their own role in effecting change. Beliefs about professional responsibility were mentioned, with students appreciating elements of the programme that focused on the societal impact of language tests. For example, Student_11 (C2) noted: 'Understanding the depth of responsibility or the weight of responsibility on the test constructor ... that changed my perspective a lot.' Reflecting on attitudes to language testing and comparing them with those of peers from different contexts was an enriching experience which also informed the students' practice.

Overall, the three main themes across the students' interviews – increased awareness of principles of language testing; better understanding of the research process; and changes that occurred in beliefs and attitudes, especially about ethics and professional roles – demonstrate a connection between the original design decisions and the programme's impact. Over the years, the MALTD also appeared to offer a pathway to further professional development for students in three distinct ways: establishing (peer) collaborations, continuing academic studies, and progressing professionally. These are exemplified in the three following sections.

Establishing (peer) collaborations

As noted previously by academic staff, the collaborative style of learning coupled with peers' diverse professional backgrounds meant that students valued opportunities to share reflections and compare practice. This learning approach has subsequently contributed to students' identification with a larger language testing community, extending beyond the MALTD through knowledge exchange.

Student_11 (C2): I have moved from being a novice into part of a discourse community, and I think that was really valuable ... It's not just about launching your own individual development, but also broadening your network of contacts, and whether that's for your own professional opportunities or whether it's to have that network to call upon for particular projects.

Feeling part of a wider community is an important legacy of the programme:

Student_5 (C7): From my cohort, I'm still in touch with some people ... who I see at conferences et cetera. And then from other cohorts, I have collaborated with other students.

Continuing academic studies

Some students enrolled on further programmes after the MALTD, including PhD studies in language testing. Consequently, their professional development involved conducting advanced research and contributing to the field through scholarly publications. For example, Student_6 (C10) reported: 'It was on my radar to do a PhD in general. I didn't know what I wanted to do and then the [MALTD] programme definitely pushed me into it.'

Progressing professionally

All students highlighted the continuing impact of the MA learning experience on their work. Some integrated the knowledge and skills acquired through the programme into their local contexts. For example, they initiated, or contributed to, teaching and training programmes in their own countries to share their experience with colleagues. Most mentioned a promotion or change of job after graduation, often involving more responsibility, as well as collaborations with different stakeholders – international partners and research associations – and, in one case, the setting up of a special interest group in language testing in their own country. Generally, the MA learning

experience was seen as helpful in linking research and practice. Sometimes, knowledge gained through the programme emphasised tensions in language assessment practice:

Student_1 (C8): I had this gut feeling that the company I was working for, well, it wasn't really operating in a way that they should have been. And I think the programme confirmed that and, well, it was so dramatic in fact, that shortly after finishing the MA, in fact, no, even before finishing, before I graduated, I actually resigned ... I realised ... the importance of validity and testing ... the construct of language testing [which] was explained to me. I actually sat up and realised the shortcomings of the instrument we were using and the methods we were using [at work].

This extract illustrates how the technical knowledge and skills gained during the MALTD may not always be sufficient for a student to instigate change in a particular institutional space. The need for an understanding of local/institutional policies and politics, and how to enact reforms in particular contexts, are aspects of practice that the MALTD curriculum might address more overtly. This issue has been recognised by current module teachers as well.

Luke Harding: As language testing professionals, we say we need to understand better how to help students to have that impact, you know, make that change ... But how do we help students to make that real change? In their context? That's a whole other step which involves all sorts of policy and political and institutional kind of knowledge ... Language testing is very embedded within policy environments, political, social, political forces.

Practical examples and further support from peers and professional associations might help to link theory and practice in contexts where established views are difficult to challenge. Case studies from different contexts were proposed by students as a desirable addition to the MALTD curriculum. When asked what further knowledge they would like to gain, former students showed interest in learning how best to share their knowledge with their own students and colleagues. For instance:

Student_14 (C2): I would like to see an activity which, for example, we didn't have back then, [on] how do you create an item writer course? ... And not just to review and develop items, but review an item writer training course.

Such suggestions reflect student awareness of the importance of sharing experience through a cascaded approach, drawing on their knowledge to design further training programmes in their own context.

Conclusion

In this chapter, we have drawn on a dataset of oral history interviews to explore the development and evolution of a Master's-level curriculum for language testing, and its impact on student learning and professional practice. Our analysis showed how the MALTD programme emerged at a time of change in Europe and represented a course design that was both radical in its online format and connected with pedagogical traditions that had held sway at Lancaster University for several decades. The programme evolved within its overarching structure – sometimes deliberately, in response to developments in the field and changes in staff interests, and sometimes in less predictable ways, related to changes in student cohorts and technological affordances. Former students indicated important aspects of their learning on the programme, including greater awareness of principles of good practice, increased knowledge of research skills, particularly in statistics, and an orientation towards ethics and professional responsibility. The programme had resonance for students in terms of building a community across cohorts, serving as a pathway to PhD studies, and providing more opportunities for career progression. At the same time, areas for improving the programme were noted, including a focus on managing change successfully in local contexts, and training in ways to cascade knowledge.

This historical perspective presents a programme that has been shaped by various factors: individual, institutional, sociohistorical, and technological. Its relative stability suggests that the curriculum may represent a framework of competences that could be generalisable to other Master's-level language testing courses (see Inbar-Lourie 2008), perhaps even constituting an 'international standard', although much about the MALTD is unique to its context and has been jointly constructed by participants over time. We argue that the oral history approach has been invaluable in foregrounding this complexity, and we encourage those teaching on other longstanding programmes to use oral history to record narratives, particularly with the original course designers.

References

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

- Bailey, K M and Brown, J D (1996) Language testing courses: what are they?, in Cumming, A and Berwick, R (Eds) *Validation in Language Testing*, Philadelphia: Multilingual Matters, 236–256.
- Brown, J D and Bailey, K M (2008) Language testing courses: what are they in 2007?, *Language Testing* 25 (3), 349–383.
- Fulcher, G (2012) Assessment literacy for the language classroom, *Language Assessment Quarterly* 9 (2), 113–132.
- Green, R (2013) *Statistical Analyses for Language Testers*, Basingstoke: Palgrave Macmillan.
- Green, R and Wall, D (2005) Language testing in the military: problems, politics and progress, *Language Testing* 22 (3), 379–398.
- Inbar-Lourie, O (2008) Constructing a language assessment knowledge base: a focus on language assessment courses, *Language Testing* 25 (3), 385–402.
- Jeong, H (2013) Defining assessment literacy: is it different for language testers and non-language testers?, *Language Testing* 30 (3), 345–362.
- Peniston-Bird, C M (2009) Oral history: the sound of memory, in Barber, S and Peniston-Bird, C M (Eds) *History Beyond the Text: A Student's Guide to Approaching Alternative Sources*, London: Routledge, 105–121.
- Quinn, F, Chater, A and Morrison, V (2020) An oral history of health psychology in the UK, *British Journal of Health Psychology* 25 (3), 502–518.
- Ritchie, D A (2014) *Doing Oral History* (Third edition), Oxford: Oxford University Press.
- Smith, R (2016) Building ‘applied linguistic historiography’: rationale, scope, and methods, *Applied Linguistics* 37 (1), 71–87.
- Taylor, L and Green, A (2020) The role of academic institutions in language testing research and consultancy, in Taylor, L and Saville, N (Eds) *Lessons and Legacy: A Tribute to Professor Cyril J Weir (1950–2018)*, Studies in Language Testing Volume 50, Cambridge: UCLES/Cambridge University Press, 175–206.
- Thompson, P and Bornat, J (2017) *The Voice of the Past: Oral History* (Fourth edition), Oxford: Oxford University Press.

Appendix: Sample interview questions

For initial programme designers

1. When you think back to the beginnings of the MALTD programme, what immediately comes to mind? (words, images, people)

Origins

2. The MALTD programme was set up in the mid-2000s. Why was it developed in that particular time period? What was happening in the department at the time that prompted a new Master's programme? Do you know of any similar or closely related programmes to this one at the time?
3. What challenges did you come across while setting up the programme?
4. What opportunities came up while setting up the programme?

Student profile

5. The original British Council students were presumably all in the same sort of professional situation. Did you have to change the initial programme design to cater for a more diverse set of students when you started the MALTD programme?

Programme design

6. Why did you decide to make the programme part-time?
7. The programme included five modules and a dissertation. Was there a particular reason why the programme was designed in this way?

Personal connections

8. Are there aspects of the original programme that you are particularly proud of? Are there aspects of the programme that might not have stood the test of time?
9. What do you feel your personal contribution to the original programme was? Where might we find your 'signature'?
10. Is there something that sums up the MALTD programme for you (for better or worse) – an incident, a particular student, an event, some feedback?

For teachers

Changes

1. During your time working on the programme, how has the programme evolved? And why do you think it has evolved in that way? (Prompts: unit topics, textbooks, naming of modules, student needs) How involved have you been in these changes?

Impact

2. In your experience, what effect has the MALTD programme had on the students?

Personal connections

3. What effect has working on the MALTD programme had on you? (Prompts: knowledge and insights into different contexts of language testing practice)
4. Do you have a favourite aspect of the programme? What is it? Why do you like it?
5. Are there any particularly memorable experiences you would like to share?

Section 2

Scholarly Reflections

10

Sign language assessment: Common misconceptions and the need for language assessment literacy development

Franz Holzknrecht, Tobias Haug
*University of Teacher Education in Special
Needs, Zurich*

The goals of this chapter are 1) to address common misconceptions about sign languages and sign language assessment, and 2) to identify needs in language assessment literacy (LAL) development for professionals working in sign language education. The chapter focusses on:

- Misconceptions about sign languages and their effects on language acquisition for deaf children
- Challenges associated with language assessment literacy development in small linguistic communities
- Suggestions for future work to foster language assessment literacy in sign language communities

Die Ziele dieses Kapitels sind 1) die Klärung gängiger Missverständnisse über Gebärdensprachen und deren Beurteilung und 2) die Feststellung des Bedarfs an Kompetenzentwicklung zum Beurteilen von Gebärdensprachen für Fachleute, die in der Gebärdensprachlehre tätig sind. Die Schwerpunkte des Kapitels liegen auf:

- Missverständnissen über Gebärdensprachen und deren Auswirkungen auf den Spracherwerb von gehörlosen Kindern
- Herausforderungen in der Kompetenzentwicklung zum Beurteilen von Sprache in kleinen Sprachgemeinschaften
- Vorschlägen für die zukünftige Arbeit zur Förderung der Sprachbeurteilungskompetenz in Gebärdensprachgemeinschaften

Introduction

Sign language assessment is a relatively new field, and sign language professionals only started collaborating with language testing researchers about a decade ago. Despite an increasing number of studies, awareness of sign languages still seems to be lacking in the spoken language testing community, and there are still some misconceptions about sign languages and Deaf¹ communities within the language testing field. On the other hand, although language assessment literacy (LAL) has been a topic in research and practice within spoken language testing for many years, it is a new area of enquiry in sign language assessment (Boers-Visker and Hammer 2022). We will begin this reflection with a brief background section outlining the main differences between spoken and signed languages to familiarize readers with the basic concepts of sign languages. We will then turn to a discussion of common misconceptions about sign languages and sign language testing, as it is important to address these for a language testing audience working predominantly in spoken languages. Finally, the third and main part of the reflection will address LAL development needs for sign language professionals and will suggest future work in this area.

Background

The *World Federation of the Deaf* estimates that there are more than 200 sign languages used by around 70 million deaf people worldwide (World Federation of the Deaf 2016). In addition, sign languages are used by (hearing) family members of the deaf, sign language interpreters, and sign language researchers. Sign languages differ from spoken languages in many respects, most obviously in the modality of language production and reception. While sign language users employ manual components (i.e., hand movement, handshape, hand orientation, place of hand-articulation) as well as non-manual components (i.e., facial expressions, eye gaze, head position and movement, torso movement) to produce language, users of spoken languages express themselves primarily through speaking and writing (although it should be acknowledged that the role of non-verbal behaviour such as gestures, eye gaze, head movement etc. is increasingly well understood as a key component of interactional competence among spoken language users). Accordingly, sign language reception depends

1 The term 'Deaf' (capitalised) is used when referring to sociocultural entities such as 'Deaf community', whereas the more inclusive 'deaf' (not capitalised) refers to individuals rather than groups to account for the increasing diversity of identities and language practices of people who are deaf or hard-of-hearing (Kusters, O'Brian and De Meulder 2017).

solely on the visual channel instead of being processed auditorily or through reading. In addition, sign languages are not used across countries to the same extent as spoken languages (e.g., English). Many countries have one or more distinct sign languages with their own unique grammar and lexicon. In Switzerland, for example, deaf people use three different sign languages, and all three languages contain multiple dialects (Boyes Braem, Haug and Shores 2012).

Misconceptions about sign languages and sign language assessment

Although awareness of sign languages has increased over the last decades and many national sign languages have obtained the same legal status as spoken languages, misconceptions about the Deaf community and about sign languages still exist (Hall, Hall and Caselli 2019). These misconceptions are often rooted in the misguided notion that sign languages may not be equivalent to spoken languages in terms of communicative richness and complexity. Such misunderstandings can have detrimental effects on children acquiring a language.

In most industrialised countries newborns undergo an early hearing screening. In the case of a confirmed diagnosis of hearing impairment, the consultation in hospitals often reflects a particular stance on advice to parents regarding language access and learning for their deaf child, namely, to expose infants to a spoken language as early as possible. While hearing children acquire a language through spoken interaction with their parents or caregivers, the situation of early language access is more complex for deaf children. Most deaf children (approximately 90%, see Mitchell and Karchmer 2004) are born into hearing families who have no knowledge of sign languages. Consequently, hearing parents of deaf infants often opt for cochlear implants to train their children in ‘hearing’ and understanding a spoken language (Humphries et al 2016), rather than exposing them to a sign language. However, as cochlear implant therapy is not fully reliable, this policy risks depriving young children of crucial language input in their first years of life, which in turn can lead to lifelong mental health difficulties (Hall 2017). Being exposed to a fully accessible language such as a sign language for deaf infants is also a crucial variable for later success in school. Researchers in deaf education and sign language linguistics thus recommend that all deaf infants acquire a sign language, both before and during cochlear implant therapy (should parents choose that option), to ensure optimal cognitive development (Amraei, Amirjalali and Ajalloueyan 2017).

Misconceptions about sign languages also still exist within the language testing community, where sign language assessment is sometimes broadly

subsumed under the category of test accommodations. Test accommodations are typically understood as special procedures or intentional changes put in place for learners with special needs, for example catering to specific requirements such as candidates with visual impairments or a nursing parent and baby. As such, students who are deaf or hard of hearing may require accommodations when they take a standardized test of a *spoken* language; however, we would like to stress that the testing of *sign* languages does not fall into the same category. Rather, sign language tests are simply tests in a different language, and sign language test-takers may require accommodations in their own right. For example, a nursing parent would require the same accommodations in tests of any language, regardless of whether the language is spoken or signed.

A likely explanation for this misconception in the language testing field is that research on sign language assessment is still sparse. Compared to the testing of spoken languages, sign language testing is a very young discipline, nested within deaf education, sign language linguistics, and applied linguistics. As outlined in the introduction, sign language assessment researchers and test developers have only relatively recently, i.e., within the last 10 years or so, started to publish their work in peer-reviewed journals outside their own field (e.g., the first publication on sign language assessment in *Language Testing* was published in 2011). Most publications on sign language assessment focus on the development and validation of sign language tests (e.g., Haug et al 2020), provide an overview of existing tests (e.g. Haug 2005), or describe methodological steps and procedures for sign language test development (e.g., Haug and Mann 2007, Haug 2011). But publications on, for example, rater behavior, standard setting, or LAL are only just beginning to emerge in the sign language testing community.

Language assessment literacy in small linguistic communities

To our knowledge, the only publication addressing LAL in the context of sign languages to date is a recent chapter by Boers-Visker and Hammer (2022). The authors propose a framework for construct-related LAL for sign language teachers, and develop a sign language assessment literacy matrix based on five quality assessment criteria for spoken language assessment (Bachman and Palmer 1996) designed to support sign language teachers in the development of reliable and valid tests. However, there is no published research, as far as we are aware, on general LAL development needs for sign language professionals. Our review here thus also included studies on LAL needs in less commonly taught languages (LCTLs), as these show clear parallels to the needs in sign languages. Two main themes emerged

from this review: (1) the significant impact of practitioners' training, certification, and literacy in the surrounding dominant language; and (2) a lack of LAL resources. The two themes will be discussed in turn, followed by recommendations for future work in this area.

Practitioners' training, certification, and dominant language literacy

Although studies on LAL in languages other than English are rare (Malone 2016), the few that are available have highlighted the role of practitioners' training and certification, as well as their proficiency in the dominant language. For example, in addressing LAL training needs of teachers of LCTLs in the USA (e.g., Mandarin, Arabic, or Hindi), Montee, Bach, Donovan and Thompson (2013) and Nier, Donovan and Malone (2009) have found that many LCTL teachers who qualified outside the USA have different or no training backgrounds (and, consequently, no certifications) compared to teachers trained in the USA. In addition, teachers of LCTLs are often L1 speakers of languages other than English, which further impacts their access to LAL training materials (Nier et al 2009).

Similar issues emerge when looking at deaf sign language teachers. As with teachers of LCTLs in the USA, the training and certification of sign language teachers is handled very heterogeneously in different national contexts, and detailed information on the various training courses and programmes offered is often not available. The other similarity to teachers of LCTLs is the low literacy level of deaf sign language teachers in the dominant language. For example, deaf sign language teachers in German Switzerland often have low proficiency in reading and writing in German (Gutjahr 2006, Traxler 2000, Wauters 2005). Thus, even the few German publications on LAL (e.g., Dlaska and Krekeler 2009, Grotjahn and Kleppin 2015) would not be fully accessible to deaf sign language teachers, let alone materials available in English.

Lack of resources

Montee et al (2013) and Nier et al (2009) also observe a dearth of resources to foster LAL training for minority languages, which is equally the case for sign languages. Underlying this scarcity of materials is a general lack of research on LAL, both on spoken minority languages and in sign languages (e.g., for spoken languages see Mellow and Begg (2013); for sign languages see Enns et al (2016)). For the latter, various national and European projects have begun to address this by developing publicly available resources for teaching and assessment. For example, within the frame of the European Centre for Modern Languages (ECML) project 'Promoting excellence in sign language

instruction’, materials on sign language assessment were developed in English and International Sign, a pidgin sign language used internationally (ECML 2022; see also Mann, Haug, Kollien and Quinto-Pozos 2014). However, more work needs to be done.

The way forward

Considering the identified needs, we suggest two main strategies to further the development of LAL in sign languages: (1) a systematic incorporation of LAL in the training and certification of different stakeholders in sign language communities (mainly teachers and researchers, but also political decision makers) and (2) steps to mitigate the language barriers of existing LAL resources published in spoken languages (such as English).

With regard to (1), it may be useful to follow Kremmel and Harding’s (2020) nine-dimensional LAL model, in which the authors suggest tailoring the focus on different LAL dimensions to different stakeholder groups. One challenge here is that, as sign language communities are usually small, stakeholder groups are often overlapping. While it may be useful in spoken language contexts for teachers to develop sound assessment pedagogy skills, for researchers to be proficient in statistical methods to analyze test data, and for test developers to be knowledgeable about language testing principles (Kremmel and Harding 2020), in sign language communities, teachers, researchers, and test developers are often the same individuals. Thus, on a local level, it is necessary to first define the different sign language stakeholder groups. An ensuing step would be to evaluate LAL training needs among these groups, such as professional organizations and training institutions of sign language teachers. Along with existing resources, this can serve as a basis for developing tailored LAL training courses.

In relation to (2), the availability of freely available online LAL resources is promising (e.g., ALTE 2018, British Council 2022, Malone 2013). An important next step would be to make more of these kinds of resources accessible and develop training opportunities for stakeholders who have insufficient command of English. LAL materials should ideally be translated into the stakeholders’ national or local sign language, or accessible in International Sign as was done in the ‘ProSign’ project (see ECML 2015). Other examples of making material accessible for deaf sign language professionals is an EALTA-funded workshop on CEFR-linked assessment in sign language education (Haug and Van den Bogaerde 2017); the development of a training course on language testing principles, test administration, and rating of a sign language interview in German Switzerland (ongoing); or the annual webinars of the EALTA SIG for SIGN on various issues in sign language assessment.

On a more general level, it seems important to continue raising awareness of sign languages and the sign language community, both within the language testing field and in society more broadly. Within the former, this would help avoid common misconceptions and may lead to increased collaboration between spoken and signed language assessment researchers (see, for example, Haug, Mann and Knoch (Eds) 2022). In the latter, the recognition of sign languages as official national languages would foster inclusivity by providing easier access to sign language translations (e.g., in healthcare settings), which would improve the lives of millions of people worldwide.

References

- ALTE (Association of Language Testers in Europe) (2018) *Guidelines for the Development of Language for Specific Purposes Tests*, available online: www.alte.org/resources/Documents/6093%20LSP%20Supplement%20-%20WEB.pdf
- Amraei, K, Amirjalali, S and Ajalloueyan, M (2017) Comparison of intelligence quotients of first- and second-generation deaf children with cochlear implants, *International Journal of Pediatric Otorhinolaryngology* 92, 167–170.
- Bachman, L F and Palmer, A S (1996) *Language Testing in Practice*, Oxford: Oxford University Press.
- Boers-Visker, E and Hammer, A (2022) Language assessment literacy in second signed language assessment contexts, in Haug, T, Mann, W and Knoch, U (Eds) *The Handbook of Language Assessment Across Modalities*, Oxford: Oxford University Press, 383–394.
- Boyes Braem, P, Haug, T and Shores, P (2012) Gebärdenspracharbeit in der Schweiz: Rückblick und Ausblick, *Das Zeichen* 90, 58–74.
- British Council (2022) *How language assessment works project*, available online: www.britishcouncil.org/exam/aptis/research/assessment-literacy
- Blaska, A and Krekeler, C (2009) *Sprachtests: Leistungsbeurteilungen im Fremdsprachunterricht evaluieren und verbessern [Language tests: Evaluate and improve performance assessment in foreign language teaching]*, Baltmannsweiler: Schneider Verlag Hohengehren GmbH.
- ECML (European Centre for Modern Languages of the Council of Europe) (2015) *Sign languages and the Common European Framework of Reference for Languages: descriptors and approaches to assessment*, available online: www.ecml.at/ECML-Programme/Programme2012-2015/ProSign/tabid/1752/Default.aspx
- ECML (European Centre for Modern Languages of the Council of Europe) (2022) *Promoting excellence in sign language instruction*, available online: www.ecml.at/Thematicareas/SignedLanguages/ProSign/tabid/4273/language/en-GB/Default.aspx
- Enns, C, Haug, T, Herman, R, Hoffmeister, R J, Mann, W and Mcquarrie, L (2016) Exploring signed language assessment tools in Europe and North America, in Marschark, M, Lampropoulou, V and Skordilis, E K (Eds) *Diversity in Deaf Education*, Oxford/New York: Oxford University Press, 171–218.
- Grotjahn, R and Kleppin, K (2015) *Prüfen, Testen, Evaluieren [Check, test, evaluate]*, München: Klett-Langenscheidt.

- Gutjahr, A (2006) *Lesekompetenz Gehörloser: Ein Forschungsüberblick [Reading competence of deaf people: A research overview]*, unpublished master's thesis, University of Hamburg.
- Hall, W C (2017) What you don't know can hurt you: The risk of language deprivation by impairing sign language development in deaf children, *Maternal and Child Health Journal* 21 (5), 961–965.
- Hall, M L, Hall, W C and Caselli, N K (2019) Deaf children need language, not (just) speech, *First Language* 39 (4), 367–395.
- Haug, T (2005) Review of sign language assessment instruments, *Sign Language & Linguistics* 8 (1/2), 61–98.
- Haug, T (2011) Methodological and theoretical issues in the adaptation of sign language tests: An example from the adaptation of a test to German Sign Language, *Language Testing* 29 (2), 181–201.
- Haug, T and Mann, W (2007) Adapting Tests of Sign Language Assessment for Other Sign Languages—A Review of Linguistic, Cultural, and Psychometric Problems, *Journal of Deaf Studies and Deaf Education* 13 (1), 138–147.
- Haug, T and Van den Bogaerde, B (2017) *Testing and assessment in classroom environment, from the first experiences to CEFR: Introduction*, EALTA workshop: Assessing sign language in the clinical and classroom environment, Learning from experience exchange, ISSR, Rome, November.
- Haug, T, Mann, W and Knoch, U (Eds) (2022) *The Handbook of Language Assessment Across Modalities*, Oxford: Oxford University Press.
- Haug, T, Batty, A O, Venetz, M, Notter, C, Girard-Groeber, S, Knoch, U and Audeoud, M (2020) Validity evidence for a sentence repetition test of Swiss German Sign Language, *Language Testing* 37 (3), 412–434.
- Humphries, T, Kushalnagar, P, Mathur, G, Napoli, D J, Padden, C, Rathmann, C and Smith, S (2016) Language choices for deaf infants: Advice for parents regarding sign languages, *Clinical Pediatrics* 55 (6), 513–517.
- 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.
- Kusters, A, O'Brien, D and De Meulder, M (2017) *Innovations in Deaf Studies: Critically Mapping the Field*, Oxford: Oxford University Press.
- Malone, M E (2013) The essentials of assessment literacy: Contrasts between testers and users, *Language Testing* 30 (3), 329–344.
- Malone, M E (2016) *Expanding understanding of language assessment literacy: Including students*, paper presented at the Language Assessment Literacy Symposium, Lancaster, September.
- Mann, W, Haug, T, Kollien, S and Quinto-Pozos, D (2014) Teaching signed languages, in Chapelle, C (Ed) *The Encyclopedia of Applied Linguistics*, Hoboken: John Wiley & Sons, Inc., 1–8.
- Mellow, J D and Begg, K (2013) Assessing North American indigenous languages, in Kunnan, A J (Ed) *The Companion to Language Assessment*, Hoboken: John Wiley & Sons, Inc., 1,759–1,768.
- Mitchell, R E and Karchmer, M A (2004) Chasing the Mythical Ten Percent: Parental Hearing Status of Deaf and Hard of Hearing Students in the United States, *Sign Language Studies* 4 (2), 138–163.
- Montee, M, Bach, A, Donovan, A and Thompson, L (2013) LCTL teachers' assessment knowledge and practices: An exploratory study, *Journal of the National Council of Less Commonly Taught Languages* 13, 1–31.

- Nier, V, Donovan, A and Malone, M E (2009) Increasing assessment literacy among LCTL instructors through blended learning, *Journal of the National Council of Less Commonly Taught Languages* 9, 105–136.
- Traxler, C B (2000) The Stanford Achievement Test, 9th edition: National norming and performance standards for deaf and hard-of-hearing students, *Journal of Deaf Studies and Deaf Education* 5 (4), 337–348.
- Wauters, L (2005) *Reading comprehension in deaf children: The impact of the mode of acquisition of word meanings*, unpublished doctoral dissertation, Radboud University, Nijmegen.
- World Federation of the Deaf (2016) *Our Work*, available online: wfdeaf.org/our-work/

11

A conceptual framework to contextualise local language assessment literacy

Newton Paulo Monteiro

Alfredo Nasser University Centre, Brazil

In this chapter I describe a conceptual framework to reflect on local professional experiences in developing language assessment literacy (LAL). My main arguments are:

- LAL should be seen as contextualised practices related to local professional roles (PRs)
- Contextualised language assessment literacy (CLAL) is developed under the influence of drivers and constraints as stakeholders take on PRs
- Reflecting on PRs, drivers, constraints and CLAL can be useful to elucidate issues impacting assessment contexts

Neste capítulo, descrevo um quadro conceitual para elaborar reflexões sobre experiências profissionais locais e sua relação com o desenvolvimento do letramento em avaliação de línguas em contexto. O argumento se desenvolve da seguinte forma:

- O letramento em avaliação de línguas deve ser entendido como um conjunto de práticas contextualizadas relacionadas aos *papéis profissionais* desempenhados localmente
- O letramento em avaliação de línguas se desenvolve em um contexto de *fatores de influência* e *fatores de restrição* à medida que os sujeitos desempenham seus papéis profissionais
- A reflexão sobre os papéis profissionais, os fatores de influência e de restrição e as condições de desenvolvimento do letramento em avaliação de línguas pode contribuir para elucidar as questões que impactam os contextos de avaliação

Introduction

In *An Introduction to Applied Linguistics*, Alan Davies (2007:3) argued that most applied linguists are engaged with language teaching, but some ‘move on to research, administration and so on’. This is a clear description of my teaching experiences and current administrative roles, alongside my engagement with language and educational assessment projects.

In addition to academic studies and resources, my professional demands have led me to reflect about the role of context in assessment literacy. In this chapter, I will refer to *contextualised language assessment literacy* (CLAL) as the assessment knowledge and skills derived from local professional experiences. Context-embedded practice fosters skills acquisition as professionals deal with the favourable forces that influence assessment literacy, such as official requirements, and the limiting ones, such as lack of resources. I will refer to the former as *drivers* and to the latter as *constraints*.

An example of a driver is the *mandate* of an assessment. Davidson and Lynch (2002:77) use the term to describe ‘the combination of forces which help to decide what will be tested and to shape the actual content of a test’. For instance, when an educational official requires a new assessment project, they may issue a mandate and assign teachers the role of assessment staff. This mandate is an external driver since it opens the opportunities for teachers to take on *professional roles* (PRs) – or functions in assessment projects – and start actions that promote their CLAL. However, the official may try to interfere in the staff’s activities. They may impose restrictions on budgets and resources, set unreasonable deadlines or try to influence the assessment content even though they are not language professionals or assessment knowledgeable. These external constraints can impact the fulfilment of teachers’ PRs and the process of CLAL acquisition. Therefore, drivers and constraints interact to produce unique professional experiences that highlight challenges in need of answers.

External drivers and *external constraints* are the forces delineating the practice of specific language assessment professionals. These forces emerge in macro (society, culture), meso (institutional) and micro contexts (staff/individuals). For example, a national educational policy (macro context) can define what assessment projects an assessment staff should develop. In addition, institutional decisions (meso context) can impact the resources available for such projects. Finally, individual views about how to conduct these projects (micro context) can influence their development and management.

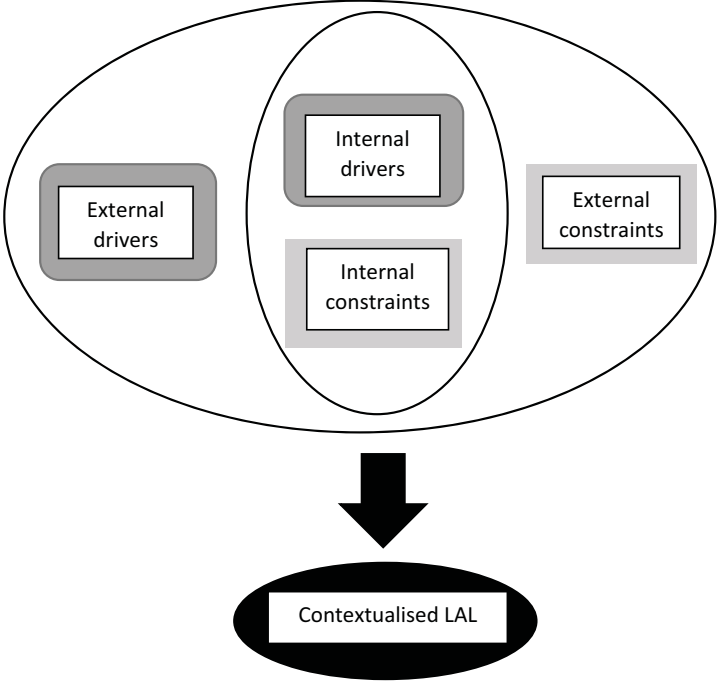
In contrast, *internal drivers* and *internal constraints* refer to favourable and limiting influences, respectively, within an assessment professional. They may entail characteristics from a person’s affective or cognitive domain. For

example, personality traits, level of knowledge and skills, work and learning styles, learning needs, attitudes, motivation, willingness to cooperate, resilience, perseverance and self-awareness can be considered drivers or constraints. Such forces can also refer to issues in the privacy or subjectivity of a person – professional experience, career aims and subjective ways of assimilating cultural references also fall within the limits of internal drivers and constraints.

PRs also vary considerably – ‘language testers often wear many hats, especially in local testing contexts’ (Dimova, Yan and Ginther 2022:348). Some stakeholders may take on different roles in every new project or multiple roles in a single project. I have been the test developer and test administrator in one case and the test developer and test manager in another, while some colleagues were in other positions.

Figure 1 illustrates this conceptual framework. Internal drivers and constraints in interaction characterise individuals in professional roles. In turn, external drivers and constraints act on individuals, interacting with the internal drivers and constraints. The whole set of factors influences the development of CLAL.

Figure 1 A conceptual framework to contextualise language assessment literacy



This framework should help generate questions for reflection and research on assessment literacy for a particular moment and context. I propose the following questions to guide my reflection in this chapter:

- What are my PRs?
- What external and internal drivers influence me?
- What external and internal constraints influence me?

The choice of these questions is justified by the fact that professionals may not be aware of all their roles and the forces influencing them. For instance, in my context I have been required to conduct projects in which I was supposed to select assessment staff. It was not clear who would be assigned which roles, including the ones I would take on. This situation may be very frequent in local contexts where the limits of roles can be blurred or the forces influencing individuals may be unclear. A framework may elucidate the invisible and unspoken factors impacting stakeholders. Assessment professionals using this framework may find out that their own contexts reveal a different set of roles, drivers and constraints. In the following sections I will illustrate how to use the conceptual framework by reflecting on my own professional history.

Mandates as external drivers of professional roles

As mentioned before, a mandate consists of ‘the combination of forces’ that leads to decisions to start an assessment project (Davidson and Lynch 2002:77). These forces include – but are not limited to – official requests. Davidson and Lynch argue that mandates can come about by people inside or outside the educational setting. This means that mandates result from institutional (meso context) or educational or economical changes (macro context). In my context*, such mandates have emerged in different moments resulting in professional roles at the medical school, the evaluation committee and the language teacher education program.

At the medical school, there have been two important mandates. The first was in 2013, when the school started its activities and issued a mandate for an admission exam. I was invited to write English items and to be an invigilator for this exam. The second mandate was issued by the Academic Dean in 2016. It required the development of an academic reading English test to exempt students from classes in English for Specific/Academic Purposes (ESP/EAP). Since I was the English teacher for these classes, I was assigned to conduct the whole process. As a result, I needed to invite colleagues to be part of

* I refer to Alfredo Nasser University Centre, a middle-size higher education institution in mid-west Brazil. A university centre is a category between college and university in the Brazilian higher education system.

the assessment staff as well as take on the roles of test manager, developer, administrator, rater and researcher.

At the evaluation committee, another mandate to revise an assessment project was issued in 2018. The evaluation committee monitors the quality of education, services, management and facilities. It also delivers an online exam with general education and Portuguese items for about 5,000 students. As the coordinator of the committee, I led a study of several problems in the exam such as issues with item quality, test management and the IT platform for online testing. A plan to address these problems was presented to the academic dean, who issued the mandate to start the test revision.

All the previous mandates were the result of specific decisions to develop a test. Another kind of mandate results from an institutional policy to select students for tutoring positions in all undergraduate programs. Tutors are high-performing students who assist other undergraduates in their learning difficulties. As of 2019, I have managed the English and Portuguese tests to select university students for these tutoring positions. Since the policy did not specify how to develop the selection exam, all the technical and management decisions were in the hands of the assessment staff. This experience of developing and administering the language tests has also contributed to my CLAL.

These four mandates have opened opportunities to take on several PRs. In addition, because I had to understand assessment principles better, I studied the specialised literature more. I also started to share what I learned with colleagues and students in classes and training sessions. This led to improvement of my knowledge and skills and fostered my reputation as assessment knowledgeable. Therefore, the mandates were drivers of my professional experiences and CLAL.

From the previous exposition, it is possible to answer the first question of the conceptual framework, *What are my PRs?*, as follows:

- Invigilator
- Item writer
- Test developer
- Test manager
- Test administrator
- Scorer/Rater
- Researcher
- Language teacher/practitioner
- (Language) teacher educator

So far, I have presented mandates as the ‘starting point[s]’ and drivers of roles in assessment projects (Davidson and Lynch 2002:77). Next I will discuss other influences on the acquisition of CLAL and project management.

Drivers and the development of CLAL

External and internal drivers are the favourable forces that incite action and promote CLAL. In Table 1, I present the external and internal drivers of my professional experiences. It is noteworthy that the specific drivers may be different for readers using the framework in their own contexts. External drivers include both macro and meso contexts, while internal drivers relate to the micro context.

Table 1 External and internal drivers that influence me

External drivers	Internal drivers
Mandates	Motivation to pursue new career development aims
Official requirements	Perseverance
Emergence of standardised tests	Sense of achievement
Professional demands	
Expertise needs in context	

Mandates and *official requirements* are external drivers closely associated with power. Both the medical school admission exam and the academic reading English test were established due to demands of the Brazilian educational law. As for the *emergence of standardised tests*, the perceived need to prepare students for national exams was the driver that led the institution to issue mandates to run an online general education exam. These drivers resulted in assessment projects and opportunities for my participation through the PRs discussed earlier.

Professional demands were another driver of my CLAL. As an ESP/EAP teacher at the medical school and language teacher educator at the language programme, I had to improve my assessment skills and know-how. This led to a search for the specialised literature and professional development courses in language assessment.

At the same time, the *expertise needs* of colleagues and academic programs became more evident. One might consider this as a limitation to professional development since specialised knowledge may be shallow in such a context. However, I took this situation as an opportunity to develop myself and contribute by tackling assessment issues. Thus, expertise needs in context became a driver for my pursuit of CLAL.

As I identified these external drivers, an awareness of my internal drivers started to emerge. Internal drivers are a set of very subjective, motivating forces that lead to action. As I conducted my reflection, I found three internal drivers that have added to my CLAL. First, my *motivation to pursue new career development aims* became a driver of professional decisions such as being more knowledgeable of assessment, keeping abreast of the field and networking with the language assessment and testing community. Along the way, however, I also faced challenges, such as the constraints that will be

discussed in the next section, which has demanded *perseverance* to attain my learning goals. Yet, whenever I could apply my newly acquired knowledge and skills in local projects and in delivering training sessions, I felt a *sense of achievement* that helped me to carry on.

Constraints and the development of CLAL

Constraints are the limiting forces that influence CLAL, such as challenges, interferences and obstacles in assessment projects. Like drivers, specific constraints might be different for readers using this framework in their contexts. In Table 2, I summarise the constraints in my professional context.

Table 2 External and internal constraints that influence me

External constraints	Internal constraints
Interference from outside the assessment staff	Lack of confidence
Miscomprehension of assessment procedures	Limited knowledge and
Limited resources, technology and staff	experience

An example of an external constraint is the *interference from outside the assessment staff*. Such interference may come from officials, the calendar and even from students in search of solutions to their concerns. For example, in some projects I was able to develop test specifications, but pressure to launch the test made it impossible to conduct any pre-operational studies. In such cases, professionals can lose an opportunity to deliver a full service and develop their CLAL by enhancing their experience.

Miscomprehension of assessment procedures is a constraint usually found among teachers, item writers and management officials. In the online exam, my staff have confirmed the usefulness of written guidelines, training sessions and systematic revision. Yet these initiatives must be kept constant, since teachers frequently forget item writing instructions. In the case of officials, like heads of department and program coordinators, it can be more of a challenge to demonstrate the soundness of assessment principles, such as developing test specifications or piloting. Assessment literacy becomes ingrained only when its principles mature.

Some external constraints are a matter of *limited resources, technology and staff*. I can count on the assistance of one or two colleagues to deliver the tests for tutor selection and on five people for the test of academic reading. The same few people write, administer and rate tests. For these projects, cooperation compensates for this limitation. Somewhat differently, the general education online exam, delivered to 5,000 students, requires a more rational division of tasks. We have a staff of 23 professionals for all phases

of the exam, which include item writing, revision and IT platform operation. Still, occasional student complaints about the visualisation of questions on screen, among other problems, reminds the staff of our resource constraints. Fostering good relations with the Technology Department has been essential to obtain assistance. CLAL implies understanding that ‘the quality of interpersonal relations ... can also impact a project’s success’ (Janssen 2022:4).

Internal constraints are another important influence on CLAL. In some situations, I *lacked confidence* about project results and my abilities to conduct them. The test of academic reading, for example, was developed under strict time pressure and I had *limited knowledge and experience*. Clearly, I could only fulfill my PRs by enhancing my learning opportunities. However, as stated by Dimova et al (2022:348), in local contexts, assessment is ‘only a portion of [one’s] responsibilities’. To cope with this dilemma, I needed *self-organisation* to achieve balance.

According to some authors (Guerin and Kunkle 2004, Songhori, Ghonsooly and Afraz 2018), self-organisation is a characteristic of natural and human systems that facilitates adaptation to a complex environment by introducing control over conflicting forces. Applied to a single professional, self-organisation can be understood as the ability to manage professional demands. Self-organisation helped me to develop an assessment-related professional identity by finding ways to transfer my assessment knowledge to diverse situations, like teaching, lecturing, conducting institutional surveys or delivering training sessions. So, even though I had multiple roles, I managed to keep some coherence between them. In turn, my peers and academic officials started to look for my skills in solving assessment issues. This approach was useful to keep focus and manage time, workload, and limited knowledge and experience more efficiently.

Conclusions

Research in language testing and assessment has emphasised the importance of local practices and collaborative work in developing assessment literacy (Baker and Riches 2018, Yan and Fan 2020). The conceptual framework presented in this chapter is a means to contextualise such local and collaborative efforts by unveiling the interaction of drivers and constraints that influence the conducting of assessment projects.

This does not mean that formal training and specialised knowledge should be put aside. Instead, contextualising LAL is a way to acquire a better comprehension of what specialised recommendations are practical for the assessment projects of a given context (Bachman and Palmer 1996). Even more, contextualising LAL may reveal the tacit knowledge available in local contexts and that is hardly considered in the field literature. This

is because the interaction between drivers and constraints can produce challenges that – once overcome – result in learning.

The CLAL framework can be a supportive tool for individuals and groups of assessment professionals to contextualise their practices. Local stakeholders can identify their roles, and the influencing drivers and constraints in individual or group reflection exercises. The framework might also be adapted to projects in qualitative research. Comparative studies of field notes, diaries and interviews might be useful to validate perceptions of drivers and constraints in context.

More in-depth studies might analyse systematically the interactions between drivers and constraints or between their external and internal versions. The relation between forces from the same category could also be studied (two external drivers acting on an individual, for example). The same PRs in different projects could be compared to elucidate how context defines them. Power relations, external interference and knowledge transfer within the team could be approached as well. In contexts where local and professional language testers work together, analyses of their relations could yield interesting insights for a better comprehension of CLAL.

These suggestions could put the CLAL conceptual framework to the test and clarify its strengths and limitations. By demonstrating its use, I hope this study has contributed to the discussions of language assessment literacy.

References

- Bachman, L B and Palmer, A S (1996) *Language Testing in Practice*, Oxford: Oxford University Press.
- Baker, B and Riches, C (2018) The development of EFL examinations in Haiti: Collaboration and language assessment literacy development, *Language testing* 35 (4), 557–581.
- Davidson, F and Lynch, B K (2002) *Testcraft: A Teacher's Guide to Writing and Using Language Testing Specifications*, Connecticut: Yale University Press.
- Davies, A (2007) *An Introduction to Applied Linguistics*, Edinburgh: Edinburgh University Press.
- Dimova, S, Yan, X and Ginther, A (2022) Local tests, local contexts, *Language Testing* 39 (3), 341–354.
- Guerin, S and Kunkle, D (2004) Emergence of constraint in self-organizing systems, *Nonlinear Dynamics, Psychology, and Life Sciences* 8 (2), 131–146.
- Janssen, G (2022) Local placement test retrofit and building language assessment literacy with teacher stakeholders: a case study from Colombia, *Language Testing Special Issue*, 1–24.
- Songhori, M A, Ghonsooly, B and Afraz, S (2018) Language teacher immunity among Iranian EFL teachers: A self-organization perspective, *Iranian Journal of English for Academic Purposes* 7 (1), 128–143.
- Yan, X and Fan, J (2020) “Am I qualified to be a language tester?”: Understanding the development of language assessment literacy among three stakeholder groups, *Language Testing* 38 (2), 219–246.

12

Practical but not atheoretical: A reflection on the promise of concept-based instruction in teacher assessment literacy development

Xuan Minh Ngo

*International Education Institute, University of
St Andrews, UK*

This chapter details the author's struggle in his attempt to make language assessment courses more practical without sacrificing their theoretical rigour, and demonstrates that:

- Teacher assessment literacy development can be seen as a process of concept formation
- This concept formation process involves the synergy of empirical and scientific concepts
- A potential approach to language assessment education is concept-based instruction (CBI)

Chương sách mô tả chi tiết những khó khăn mà tác giả gặp phải trong quá trình làm cho các khóa học kiểm tra đánh giá ngôn ngữ trở nên thực tế hơn mà không làm mất đi tính chặt chẽ về mặt lý thuyết. Chương sách chứng minh rằng:

- Quá trình phát triển năng lực kiểm tra đánh giá của giáo viên có thể được xem như quá trình kiến tạo các khái niệm kiểm tra đánh giá.
- Quá trình kiến tạo các khái niệm kiểm tra đánh giá này có sự kết hợp giữa các khái niệm thực tiễn và khoa học.
- Một cách tiếp cận hiệu quả để phát triển năng lực kiểm tra đánh giá của giáo viên là hướng dẫn dựa trên khái niệm, một phương pháp sư phạm dựa trên học thuyết văn hóa xã hội của Vygotsky

Introduction

I often think of myself as a lucky test-taker. As a secondary school student in Vietnam, I was fortunate enough to win top awards from many English competitions at the city, provincial, and national levels. My seemingly endless luck with those tests largely accounted for why I was so eager to attend a language assessment course during my bachelor's TESOL program. As a third-year student with extensive test-taking experience but no formal training about language testing, I initially found the course a truly eye-opening experience even though it covered only concepts such as validity, reliability, and different test types in a rudimentary fashion. However, the final test dealt a serious blow to my positive feelings about the course. Even with my limited understanding, I could sense something wrong when I was required to recall and write down definitions of basic concepts rather than demonstrate my ability to use them in practical assessment situations. I eventually did reasonably well on this test despite falling short of my initial high distinction target. Nevertheless, encountering a bad test in a course that was supposed to showcase the best of assessment practices somehow dampened my interest in assessment.

Fortunately, this interest was soon rekindled by my in-service experiences when immediately in my first semester as a university lecturer in Vietnam, I was assigned to design midterm and end-term speaking tests. To compensate for my lack of training in test development and item writing, I resorted to copying the task formats of well-regarded international exams and trying to create similar questions without any idea about the underlying test constructs or the test specifications. As I moved through the ranks and became a course coordinator myself, I was responsible for designing tests of all four skills and was soon faced with the challenges of writing receptive skill items. Working on a tight deadline and meagre pay (approximately two weeks and \$20 for each four-skill test), I had little choice but to take reading and listening items that best matched the syllabuses directly from commercial exam preparation materials with little modification. This copy-paste approach to test design seemed to be the norm in my country (Ngo 2018); nevertheless, deep down inside I still felt a sense of unease and dreamt of creating my own tests from scratch one day.

The chance to realise my dream came when I did my Master's at a major Australian university. I enrolled on the language testing and assessment course with high hopes, and fortunately the course did not disappoint. It both deepened and expanded my understanding of the fundamentals of assessment, covering a range of concepts from test usefulness, formative assessment, and the principles of assessing the four macro skills to issues of fairness and ethics. More importantly, the course offered me the chance to try out writing test specifications and items via a small-scale test development

project, which later turned out to be of tremendous value for my job as an item writer for a high-stakes standardised listening test in Vietnam.

Having accumulated some experience in writing both low-stakes and high-stakes tests, I was keen to spread my ‘wisdom’ and volunteered to teach language assessment courses to both pre-service and in-service teachers. Nevertheless, regardless of my utmost enthusiasm and arduous effort, my undergraduate students did not seem to grasp the assessment concepts and scored poorly on the follow-up quizzes. As for my in-service teacher trainees, I still vividly remember being interrupted by one teacher who asked: ‘Why do we have to learn about validity? We’re classroom teachers, not test designers!’ In response, I poured my heart out to convince the trainees of the importance of understanding theoretical concepts, but eventually most still expressed a rather lukewarm interest in my lecture.

These struggles as an assessment educator and my growing interest in assessment have motivated me to explore how to design effective assessment courses and ultimately pursue a PhD project into teacher assessment literacy (TAL). As I would soon discover from the TAL literature, my challenges were not so unique after all.

My conundrum with language assessment education

Recent studies such as Vogt and Tsagari (2014), Atay and Mede (2017), and Xu and Brown (2017) have found many teachers to be unprepared for their classroom assessment duties mainly because their pre-service assessment courses were too theoretical (Csépes 2014, Lam 2015). The TAL literature also drew my attention to the fact that my assessment courses did not place sufficient emphasis on alternative assessment and assessment for learning, and how to harmonise large-scale standardised testing and classroom assessment, as also pointed out in Csépes (2014) and Lam (2015). It gradually dawned on me that my experience in test development and item writing had somehow blinded me to the fact that classroom teachers spent more time on informal assessment activities such as observation and oral questioning and formative assessment activities such as giving feedback rather than writing tests. Hence, it is hardly surprising that my teacher trainees expressed only a lukewarm interest (if at all) in construct validity and item facility indices. Nevertheless, I was comforted by the knowledge that teachers, overall, remained enthusiastic about receiving assessment training if such training is of practical use (Fulcher 2012), and covers the topics closely related to their classroom practices such as alternative assessment (Vogt and Tsagari 2014) and giving feedback (especially informal, self and peer feedback) (Atay and Mede 2017).

My engagement with the TAL literature, nonetheless, left me with a conundrum: how can I reconcile my teacher trainees’ preferences with

my belief in the role of theory in quality assessment education? On the one hand, my trainees, especially in-service teachers, apparently prefer to forgo abstract knowledge in favour of what can be taken straight back to their classrooms. On the other hand, I am convinced that if I am to discard all theoretical concepts and provide only assessment tips and tricks, my (student) teachers will end up becoming ‘teaching and assessment machines’ who act on gut instinct instead of fully fledged teaching professionals who are guided by sound principles and follow evidence-based practices. Fortunately, a potential answer came as I ventured into Vygotsky’s sociocultural theory (SCT), especially his writing on concepts (Vygotsky 1986), as part of my PhD project.

Teacher assessment literacy development as concept formation

‘Practical experience also shows that direct teaching of concepts is impossible and fruitless. A teacher who tries to do this usually accomplishes nothing but empty verbalism, a parrotlike repetition of words by the child, simulating a knowledge of the corresponding concepts but actually covering up a vacuum’ (Vygotsky 1986:150). ‘This is so true!’ I said to myself as I came across this quote. Simultaneously, my mind also drifted back to the class session when I passionately talked about test usefulness. As I had little difficulty comprehending this concept during my postgraduate assessment course, I naively assumed that my undergraduate students would find it equally easy to internalise the same concept. Nevertheless, the follow-up quiz soon revealed that my heartfelt lecture was an abysmal failure with more than half of the students scoring under 50%.

Struck by Vygotsky’s comment on the ineffectiveness of the direct teaching of concepts, I devoted an entire week to reading Vygotsky (1986) and learnt that in SCT, a concept is a powerful psychological tool that can be used to regulate thinking. Vygotsky (1986) divides concepts into spontaneous/empirical/everyday concepts and scientific/academic concepts. While empirical concepts originate from everyday life, are ‘saturated with experience’ (1986:193), and can thus be easily employed in practical situations, they cannot be used ‘freely and voluntarily’ to ‘form abstractions’ (1986:148–149). By contrast, scientific concepts are the result of formal instruction, and thus systematic and widely applicable; however, they may be excessively abstract and take time to develop into tools for thinking. Interestingly, instead of putting spontaneous and scientific concepts in opposition, Vygotsky views them as in a dialectical relationship in which they are ‘related and constantly influence each other’ and constitute ‘parts of a single process: the development of concept formation’ (1986:157). To be specific, Vygotsky argues that:

The development of a spontaneous concept must have reached a certain level for the child to be able to absorb a related scientific concept. In working its slow way upward, an everyday concept clears a path for the scientific concept and its downward development ... Scientific concepts, in turn, supply structures for the upward development of the child's spontaneous concepts toward consciousness and deliberate use. (1986:194)

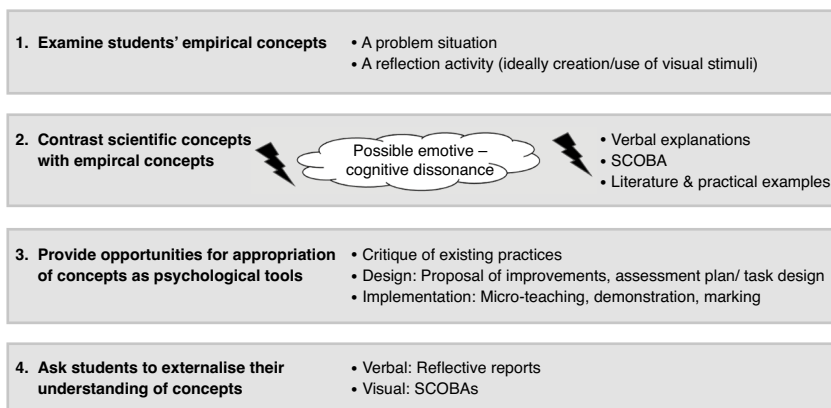
This idea marked my awakening moment. Suddenly it became crystal clear why my teaching of assessment concepts was unsuccessful: I failed to build a bridge between my students' empirical assessment concepts and the focal scientific concepts. It is true that I did ask them about their previous assessment experiences, but that activity served as a warm-up rather than constituting a central part of my lesson. As I investigated how to integrate scientific and empirical concepts in instruction to facilitate the formation of true concepts defined as internalised scientific concepts that guide teacher practices (Esteve, Farró, Rodrigo and Verdía 2021), I came across an approach known as concept-based instruction (Lantolf, Xi and Minakova 2020).

Concept-based instruction in language assessment education

Concept-based instruction (CBI) has been adopted in applied linguistics for teaching of various language features over the past two decades, yet CBI adoption in language teacher education remains rather limited (Lantolf et al 2020). Therefore, to assist with the implementation of CBI in language assessment education, I have synthesised relevant literature (Esteve Ruesca 2018, Karpov 2018, Lantolf et al 2020) to develop a figure that summarises a possible procedure for a CBI language assessment session (see Figure 1).

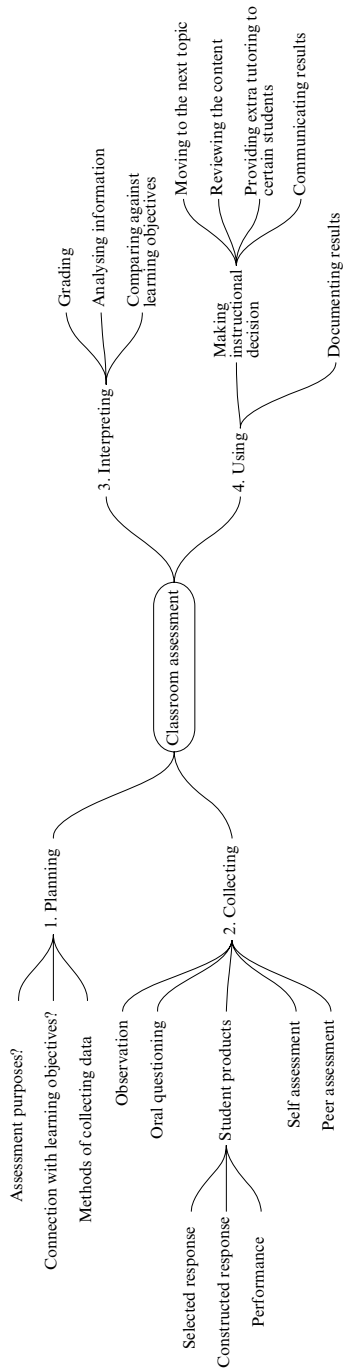
As can be seen from Figure 1, the first stage of a CBI language assessment lesson should involve students externalising their empirical concepts via a reflection activity which ideally involves a visual representation of their understanding in the form of a mind map or visual narrative (Esteve Ruesca 2018) and/or a problem situation where students are presented with questions to 'which they think they know the answer but then realise that their answers are wrong and may even contradict one another' (Karpov 2018:106). For instance, student teachers could be asked to provide examples of classroom assessment or draw a mind map to describe the process of classroom assessment. In the second stage, students should be presented with the equivalent scientific concepts initially via verbal explanations and then graphic representations known as Schemas for the Complete Orienting Basis of Action (SCOBAs) such as in Figure 2, which

Figure 1 A CBI language assessment session procedure



I developed based on accounts of classroom assessment such as Davison and Leung (2009) and Katz and Gottlieb (2013). Due to their visual nature, SCOBAs are ‘more memorable than purely verbal explanations’ (Lantolf et al 2020:330), and if SCOBAs are procedural charts, they can provide procedural knowledge of concepts, hence better enabling students to utilise these concepts as tools for thinking (Karpov 2018). While SCOBAs tend to be designed in advance by teachers, students should be engaged in modifying and expanding SCOBAs to promote concept internalisation (Esteve et al 2021, Karpov 2018). A key part of the second stage involves contrasting scientific concepts represented in the literature with empirical concepts from students’ everyday experiences. This crucial task could be done by comparing visual representations of empirical concepts outlined in the first stage with the SCOBAs representative of scientific concepts (Esteve et al 2021); and to further stimulate concept formation, students should be offered relevant academic literature and practical examples while contrasting the two types of concepts (Esteve Ruesca 2018). This contrastive analysis activity will likely lead to emotional and intellectual tensions as students’ misconceptions are confronted (Lantolf et al 2020); however, these tensions, if properly resolved under the mediation of teacher educators, are crucial for the formation of true concepts (Johnson and Golombek 2020) or TAL development in this case. Returning to the example of classroom assessment in the previous stage, student teachers may feel surprised or even confused to learn from Figure 2 that classroom assessment also involves teacher observation or oral questioning. Nevertheless, proper mediation, which may consist of teacher educators’ initial explanation and demonstration as well as subsequent accessible readings of these lesser-known forms of assessment, is likely to trigger

Figure 2 An example SCOBA on the process of classroom assessment



student teachers' interest in trying out observation and oral questioning in their (future) classrooms.

In the third stage, students should be afforded multiple opportunities to use concepts as psychological tools. For example, they could be asked to critique and propose changes to make their past classroom assessment experiences more formative. Importantly, both pre-service and in-service teachers should be given the chance to materialise the scientific concepts in classroom contexts via formulating comprehensive assessment plans that cover all four stages of classroom assessment as in Figure 2: designing assessment tasks (including classroom tests), marking writing scripts or oral performances, and implementing embedded assessment such as oral questioning and observation via micro-teaching and demonstrations. This will allow academic concepts to 'be linked in sustained, intentional, and systematic ways to the day-to-day activities of teaching/learning in classrooms' and 're-structure teachers' everyday understandings and gradually enable them to enact their conceptions and related activities of teaching purposefully' (Johnson and Golombek 2020:122). In the final phase, to consolidate the formation of true concepts, students should be asked to write reflective reports and create new SCOBAs to compare their empirical concepts in the first stage with their current (more expanded) true concepts (Esteve Ruesca 2018).

Challenges of adopting concept-based instruction

The potential of CBI in teacher education has recently been documented in Esteve Ruesca (2018) and Esteve et al (2021); nonetheless, there has been hardly any research into the use of CBI in language assessment education. This gap, to some extent, can be attributed to the challenges involved in the design and implementation of such CBI-informed courses.

The first challenge is arguably the selection of scientific concepts which serve as the minimal unit of instruction in a CBI program (Lantolf et al 2020). From the perspective of child psychology, Vygotsky argues that 'the development of a spontaneous concept must have reached a certain level for the child to be able to absorb a related scientific concept' (1986:194). Placed in the context of teacher education, it can be inferred that the selected scientific concepts should resonate with teachers' classroom assessment experiences. For this purpose, the heuristic proposed by Taylor (2013) as well as empirical research by Fulcher (2012) and Lam (2015) offer good starting points for selecting scientific assessment concepts that may be of interest to classroom teachers; nevertheless, teacher educators will likely need to conduct further research to choose relevant scientific concepts for their own programs given TAL's contextualised nature (Lam 2015, Xu and Brown 2017).

Another hurdle to CBI implementation is the design of SCOBAs, which must represent concepts 'in such a way that (student) teachers perceive

them as both useful and understandable’ (Esteve et al 2021:28). To this end, Karpov (2018) recommends using procedural SCOBAs, especially the if–then flowcharts to aid students in decision making. Nonetheless, as Karpov (2018) has acknowledged, there exist ‘ill-structured problems, that is, problems for solution of which there is no strict system of “if-then” rules’ (2018:112), in which cases, alternative forms of conceptual SCOBAs or procedural SCOBAs (Esteve et al 2021) may be preferable.

Conclusion

Regardless of the challenges, it is worth exploring CBI as a viable approach for promoting teacher assessment literacy, especially when the prevalent transmission model which focuses heavily on imparting students with theoretical and context-neutral knowledge (Esteve Ruesca 2018) has failed to produce the desired TAL levels (Atay and Mede 2017, Vogt and Tsagari 2014, Xu and Brown 2017). However, as with any innovative approach, more research is needed to ensure the successful adoption of CBI in language assessment courses. Hence, it is my ambition that this reflection will contribute to generating more interest in this promising research avenue.

References

- Atay, D and Mede, E (2017) English language teachers’ assessment literacy: The Turkish context, *Dil Dergisi* 1 (168), 43–60.
- Csépes, I (2014) Language assessment literacy in English teacher training programmers in Hungary, in Horváth, J and Medgyes, P (Eds) *Studies in Honour of Nikolov Marianne*, Pécs: Lingua Franca Csoport, 399–411.
- Davison, C and Leung, C (2009) Current issues in English language teacher-based assessment, *TESOL Quarterly* 43 (3), 393–415.
- Esteve, O, Farró, L, Rodrigo, C and Verdía, E (2021) Meaningfully designing and implementing SCOBAs in socioculturally-based L2 teacher education programs, *Language and Sociocultural Theory* 8 (1), 8–34.
- Esteve Ruesca, O (2018) Concept-based instruction in teacher education programs in Spain as illustrated by the SCOBA-mediated Barcelona formative model: Helping teachers to become transformative practitioners, in Lantolf, J P, Poehner, M E and Swain, M (Eds) *The Routledge Handbook of Sociocultural Theory and Second Language Development*, New York: Routledge, 487–504.
- Fulcher, G (2012) Assessment literacy for the language classroom, *Language Assessment Quarterly* 9 (2), 113–132.
- Johnson, K E and Golombek, P R (2020) Informing and transforming language teacher education pedagogy, *Language Teaching Research* 24 (1), 116–127.
- Karpov, Y V (2018) Acquisition of scientific concepts as the content of school instruction, in Lantolf, J P, Poehner, M E and Swain, M (Eds) *The Routledge Handbook of Sociocultural Theory and Second Language Development*, New York: Routledge, 102–116.

- Katz, A and Gottlieb, M (2013) Assessment in the classroom, in Chapelle, C A (Ed) *The Encyclopedia of Applied Linguistics*, Hoboken: Wiley Blackwell, 1–7.
- Lam, R (2015) Language assessment training in Hong Kong: Implications for language assessment literacy, *Language Testing* 32 (2), 169–197.
- Lantolf, J P, Xi, J and Minakova, V (2020) Sociocultural theory and concept-based language instruction, *Language Teaching* 54 (3), 327–342.
- Ngo, X M (2018) Sociopolitical contexts of EFL writing assessment in Vietnam: Impact of a national project, in Ruecker, T and Crusan, D (Eds) *The Politics of English Second Language Writing Assessment in Global Contexts*, New York: Routledge, 47–59.
- 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.
- Vygotsky, L S (1986) *Thought and Language*, Cambridge: MIT Press.
- Xu, Y and Brown, G T L (2017) University English teacher assessment literacy: A survey-test report from China, *Papers in Language Testing and Assessment* 6 (1), 133–158.

13 Collaborative assessment cultures and the development of LAL

Henrik Bøhn

University of South-Eastern Norway

Dina Tsagari

Oslo Metropolitan University, Norway

This chapter presents conceptions of language assessment literacy (LAL) in Norway and discusses the role of collaborative assessment cultures in the development of teacher LAL. The focus is on:

- The Norwegian educational context
- Research findings regarding assessment practices and teacher LAL in Norway
- How assessment cultures may contribute to enhancing teacher LAL in Norway and similar contexts

Introduction

Questions of what language assessment literacy (LAL) is and how it can be developed have preoccupied researchers, teachers and other stakeholders for a number of years (Berry, Sheehan and Munro 2019, Taylor 2013, Vogt and Tsagari 2014). There are also discussions regarding the extent to which the construct is impacted by contextual factors (Inbar-Lourie 2017). In this chapter we discuss the construct of teacher LAL against a Norwegian backdrop and reflect on how collaborative assessment cultures can contribute to developing it in this context and beyond.

The Norwegian educational context

The Norwegian school system is divided into primary school (Grades 1–7), lower secondary school (Grades 8–10) and upper secondary school (Grades 11–13). The subject English is compulsory from Grade 1 (age six) to Grade 11 (age 17), and instruction is curriculum-based throughout, involving a number of learning objectives. Summative assessment, in the form of marks,

is reserved for the lower and upper secondary levels and mainly given for ‘overall achievement’, awarded by the students’ subject teachers on the basis of various forms of classroom-based assessment. Additionally, all students are randomly selected to take a limited number of exams in some subjects. This system, where the occasional exam mark is awarded in addition to the overall achievement marks, can be explained historically. Since the late 19th century Norwegian teachers have had substantial autonomy, exerting considerable control over students’ school-leaving marks through the achievement-mark system. However, the additional use of exams is intended to have a calibrating effect on teachers’ grading practices and to have an external control function, monitoring learning in schools (Lundahl and Tveit 2014).

Formative assessment has been highly prioritised in Norway since the late 2000s due to the educational authorities’ realisation that such assessment may greatly improve students’ learning outcomes (Meld St 16 (2006–2007) 2007). In fact, in 2009 it was made statutory in the Regulations to the Education Act. These stipulate, for example, that pupils are to ‘participate in the assessment of their own work and reflect on their own learning and development’, ‘to understand what will be expected of them’, to ‘be informed of what they have achieved’ and to ‘obtain guidance on how they can [improve]’ (Norwegian Ministry of Education and Research 2006/2021, 3–10, our translation). In addition to the decision to give formative assessment legal status, the national authorities also initiated a number of measures to improve teachers’ assessment literacy (AL) in this area. Hence, Assessment for Learning (AfL) was made a national priority area in 2010, involving in-service training programmes, online resources for teachers and campaigns to make teachers employ assessment formatively (Norwegian Directorate for Education and Training 2019).¹

In terms of formal teaching qualifications, language teachers in Norway have diverse educational backgrounds, meaning that their language competence and exposure to assessment theory and practice can vary considerably. At the primary and lower secondary level most teachers have completed three-, four- or five-year general teacher education programmes.² However, there are also a number of teachers, particularly at the lower and upper secondary levels, who hold a bachelor’s or Master’s degree in languages, in addition to a required, one-year study of educational theory and practice. Beyond this, there are other avenues into the teaching

1 In this chapter we use the terms ‘formative assessment’ and ‘Assessment for Learning (AfL)’ synonymously.

2 The number of years depends on when they completed their education. Before 1994 this programme lasted three years, from 1994 to 2017 it was extended to four years, and after the latest national educational reform, it has been extended to a five-year Master’s programme.

profession, such as various Master's programmes in English language teacher education. Interestingly, a substantial number of teachers (14% in 2019) at the primary and lower secondary levels do not have higher education qualifications at all (Foss 2019). As for assessment training in tertiary education, this is usually provided in the different modules that the teacher trainees are required to take, but the extent of the training depends on the individual tertiary institution.

Prior to the mid-2000s few measures had been taken to hold schools and teachers to account for students' achievements (Tveit 2014). After the turn of the millennium, however, national tests were introduced to monitor students' basic skills in English, reading and numeracy at certain levels. Moreover, the revision of the Regulations to the Education Act in 2009 not only included stipulations regarding formative assessment, but also laid down explicit requirements for the documentation of student competence (Sylte 2014, Tveit 2014). In addition, the overhaul of the national curriculum in 2020 saw initiatives to introduce more psychometrically oriented test methods to strengthen the validity and reliability of exams, on the basis of recommendations made by a national committee looking into the quality of the examination system (Blömeke et al 2020).

Research on assessment practices and LAL in Norway

According to the Norwegian Directorate for Education and Training (2019), the situation in terms of teachers' *formative* assessment literacy seems to have improved after the implementation of the above-mentioned government schemes from the mid-2000s onwards. However, there is evidence that further training is needed, especially regarding the implementation of AfL in accordance with the underlying ideas of this method (Norwegian Directorate for Education and Training 2019). In addition, concerns have been made about *summative* assessment, in the sense that excessive variability has been identified in teachers' overall achievement marking and in exam results (Blömeke et al 2020). The educational authorities have therefore made it clear that measures must be taken to improve the validity and reliability of summative assessments (Norwegian Directorate for Education and Training 2020).

When it comes to Norwegian studies specifically investigating LAL, research is limited. Three exceptions are Tsagari (2021), Bøhn and Tsagari (2021) and Bøhn and Tsagari (2022). In the study on teacher LAL conducted by Tsagari (2021), focusing on the quality of EFL teachers' assessment practices and training priorities, the teachers reported being concerned with the assessment of traditional skills and components, especially speaking and writing. Teachers also confirmed that they were committed to formative

assessment, in the sense of using a number of different methods to promote students' learning and plan their own teaching.

In the study by Bøhn and Tsagari (2021), investigating teacher educators' understanding of what LAL entails, it was found that the informants thought of LAL in terms of four constructs: (i) *Disciplinary competence*, (ii) *Assessment-specific competence*, (iii) *Pedagogical competence* and (iv) *Collaboration competence*. The first construct is related to aspects such as knowledge of language theory and curriculum-related issues, including culture and literature. The second one concerns, for example, understanding the purpose of assessment, knowing what should be assessed and being able to design appropriate tasks. The third construct involves knowledge and skills relating to how learning can be promoted, knowing how to differentiate instruction, and conveying content knowledge. Finally, *Collaboration competence* was explained as the ability of teachers to work together with colleagues, school leaders and students in order to create a fruitful assessment culture. The emphasis put on this dimension by the informants was hypothesised to be influenced by the educational authorities' campaigns to encourage such collaboration to build assessment cultures among teachers (see next section, below).

In the study by Bøhn and Tsagari (2022) very similar discoveries were made. In addition to the four constructs mentioned by the teacher educators (cf. previous paragraph), the teachers also pointed to *Metacognitive skills* as important for teacher LAL. This dimension was explained as the ability of teachers to reflect on the usefulness of the assessment methods employed, thereby preventing perfunctory practices which do not contribute to valid, reliable and fair assessments.

Interestingly, many of the above-mentioned LAL aspects are in line with theoretical and empirical descriptions of LAL mentioned in the international research literature (e.g. Brindley 2001, Fulcher 2012, Kremmel and Harding 2019, Taylor 2013). However, two conspicuous differences are the very strong focus placed on formative assessment and the relatively little weight given to testing-related knowledge and skills (such as statistics and how to create multiple-choice items) by the Norwegian informants (Bøhn and Tsagari 2021, 2022).

Enhancing LAL through the development of assessment cultures

The Norwegian educational authorities' recommendation to develop assessment cultures as a means to enhancing LAL and improving quality in assessment is widely supported in the research community (Birenbaum 2014, Blömeke et al 2019, Sandvik and Fjørtoft 2016, Smith 2016) and by teachers and teacher educators, as mentioned above. Such assessment

cultures may support both summative and formative assessment practices (Blömeke et al 2020, Sandvik and Fjørtoft 2016) and find theoretical support in the ideas of situated learning and communities of practice (Lave and Wenger 1991, Wenger 1998). Broadly speaking, the development of an assessment culture involves collaboration between stakeholders to reach a common understanding of the nature of assessment, assessment purposes, and how assessment should be carried out. Regarding summative assessment, the development of an assessment culture may involve teachers coming together to discuss learning objectives, assessment criteria, scoring procedures and how to judge student performance (Bøhn 2016). As for formative assessment, it also means involving students, since formative assessment requires ‘activating students as owners of their own learning’ (Black and Wiliam 2018:560). In addition, school leaders should be included, since a proper assessment culture requires the entire school organisation to have a collective understanding of what AfL entails (Smith and Engelsen 2012). As a means to this end stakeholders should develop a ‘shared language’, which may help them make sense of what should be assessed and how (Hopfenbeck, Tolo, Florez and El Masri 2013:32, Smith 2016).

Moreover, the development of teacher LAL may be facilitated in in-service training schemes, where expert professionals can assist in providing guidance along the way (Smith 2016). However, competence in this area can also be developed without the supervision of experts, provided that a good collaborative atmosphere is promoted, and, preferably, that efforts are initiated and organised by the school leadership (Smith and Engelsen 2012). Beyond in-service training initiatives, teacher education also has an important role to play. According to Blömeke et al (2019:65), ‘teacher education institutions are by far best suited for developing good assessment competence in the areas of both formative and summative assessment for all teachers’. Hence, the development of collaborative practices to enhance LAL should be high on the agenda in teacher education.

However, one challenge in the development of formative assessment literacy is avoiding instrumental approaches built on generic ‘how-to’ checklists, which may miss out on the contextualized and complex nature of AfL and lead to superficial practices (Hopfenbeck et al 2013). Hence, teachers need, as discussed earlier in this chapter, to develop good understanding of the underlying ideas of formative assessment and to be able to tailor such assessment to the needs and characteristics of their own students and school environment (Smith 2016). This may require school leaders to encourage an atmosphere of trust and openness, and for teachers to be given the opportunity to reflect critically together on their own practices (Smith and Engelsen 2012). The latter point may be linked to the use of metacognitive skills, as mentioned above.

A host of other factors other than assessment culture may of course affect the extent to which teacher LAL will evolve: time constraints, accountability demands, resources available, conflicting interests, etc. This goes for both the Norwegian educational context and other contexts. However, given the importance of valid, reliable and fair summative assessments, the positive effects of formative assessment on student achievement (Black and Wiliam 1998, Wiliam 2018), and the impact of assessment cultures on LAL development, stakeholders in outcomes-based educational contexts where teachers are involved in summative assessments and/or formative assessment may want to consider how a fruitful assessment culture can be built in that context.

References

- Berry, V, Sheehan, S and Munro, S (2019) What does Language Assessment Literacy Mean to Teachers?, *English Language Teaching Journal* 73 (2), 113–123.
- Birenbaum, M (2014) Conceptualizing Assessment Culture in School, in Wyatt-Smith, C, Klenowski, V and Colbert, P (Eds) *Designing Assessment for Quality Learning*, Dordrecht: Springer, 285–302.
- Black, P and Wiliam, D (1998) Assessment and Classroom Learning, *Assessment in Education: Principles, Policy & Practice* 5 (1), 7–74.
- Black, P and Wiliam, D (2018) Classroom Assessment and Pedagogy, *Assessment in Education: Principles, Policy & Practice* 25 (6), 551–575.
- Blömeke, S, Skillinghaug, S, Blikstad-Balas, M, Eggen, P-O, Fjørtoft, H, Gamlem, S T M, Prøitz, T, Tveit, S, Helgesen, R, Johannessen, S, Minken, M, Waage, M and Walker, M J (2019) *Kunnskapsgrunnlag for Evaluering av Eksamensordningen [Platform of Knowledge for the Evaluation of National Exams]*, Oslo: Norwegian Directorate for Education and Training, available online: www.udir.no/tall-og-forskning/finn-forskning/rapporter/Kunnskapsgrunnlag-for-evaluering-av-eksamensordningen/#
- Blömeke, S, Skillinghaug, S, Blikstad-Balas, M, Eggen, P-O, Eira, K-I, Fjørtoft, H, Gamlem, S T M, Gilje, Ø, Tveit, S, Helgesen, R, Johannessen, S, Minken, M, Waage, M and Walker, M J (2020) *Vurderinger og Anbefalinger om Fremtidens Eksamen [Evaluations and Recommendations Concerning Future Exams]*, Oslo: Norwegian Directorate for Education and Training, available online: www.udir.no/eksamen-og-prover/eksamen/vurderinger-og-anbefalinger-fremtidens-eksamen/bakgrunn/#
- Bøhn, H (2016) *What is to be Assessed? Teachers' Understanding of Constructs in an Oral English Examination in Norway*, unpublished PhD thesis, University of Oslo.
- Bøhn, H and Tsagari, D (2021) Teacher Educators' Conceptions of Language Assessment Literacy in Norway, *Journal of Language Teaching and Research* 12 (2), 222–233.
- Bøhn, H and Tsagari, D (2022) Language Assessment Literacy: Understanding the Construct from Norwegian EFL Teachers' Perspective, *Studies in Language Assessment* 11 (1), 119–148.
- 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.
- Foss, E S (2019) *Lærere uten Pedagogisk Utdanning i Grunnskolen 2018/2019 [Teachers without Pedagogical Education in Primary and Lower Secondary School 2018/2019]*, Oslo: Statistics Norway, available online: ssb.brage.unit.no/ssb-xmlui/bitstream/handle/11250/2627226/RAPP2019-19_web.pdf?sequence=1
- Fulcher, G (2012) Assessment Literacy for the Language Classroom, *Language Assessment Quarterly* 9 (2), 113–132.
- Hopfenbeck, T, Tolo, A, Florez, T and El Masri, Y (2013) *Balancing trust and accountability? The Assessment for learning programme in Norway*, available online: www.oecd-ilibrary.org/education/balancing-trust-and-accountability-the-assessment-for-learning-programme-in-norway_5k3txnpqlsnn-en
- Inbar-Lourie, O (2017) Language Assessment Literacy, in Shohamy, E, Or, I G and May, S (Eds) *Language Testing and Assessment* (Third edition), Cham: Springer, 257–270.
- Kremmel, B and Harding, L (2019) 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.
- Lave, J and Wenger, E (1991) *Situated Learning: Legitimate Peripheral Participation*, Cambridge: Cambridge University Press.
- Lundahl, C and Tveit, S (2014) Att Legitimera Nationella Prov i Sverige och i Norge – en Fråga om Profession och Tradition [Legitimising National Tests in Sweden and Norway – a Question of Profession and Tradition], *Pedagogisk forskning i Sverige* 19 (4–5), 297–323.
- Meld St 16 (2006–2007) (2007) ... og Ingen Sto Igjen: Tidlig Innsats for Livslang Læring. [... and No One Lingered: Early Initiatives for Life-Long Learning], Oslo: Norwegian Ministry of Education and Research, available online: www.regjeringen.no/no/dokumenter/stmeld-nr-16-2006-2007-/id441395/
- Norwegian Directorate for Education and Training (2019) *Erfaringer fra Nasjonal Satsing på Vurdering for Læring (2010–2018) [Report from the Evaluation of the National Priority Area Assessment for Learning (2010–2018)]*, Oslo: Norwegian Directorate for Education and Training, available online: www.udir.no/tall-og-forskning/finn-forskning/rapporter/erfaringer-fra-nasjonal-satsing-pa-vurdering-for-laring-2010-2018/1.innledning/#1.1-bakgrunn
- Norwegian Directorate for Education and Training (2020) *Vurderinger og Anbefalinger om Fremtidens Eksamen [Evaluations of, and Recommendations for, Exams in the Future]*, Oslo: Norwegian Directorate for Education and Training, available online: www.udir.no/eksamen-og-prover/eksamen/vurderinger-og-anbefalinger-fremtidens-eksamen/
- Norwegian Ministry of Education and Research (2006/2021) *Forskrift til Opplæringslova [Regulations to the Education Act]*, available online: lovdata.no/dokument/SF/forskrift/2006-06-23-724/KAPITTEL_5#KAPITTEL_5
- Sandvik, L V and Fjørtoft, H (2016) Å Utvikle Tolkningsfelleskap i Skolen: Et Verktøy og et Kart for Vurdering [Developing Assessment Cultures in Schools: A Tool and a Map for Assessment], in Fjørtoft, H and Sandvik, L V (Eds) *Vurderingskompetanse i Skolen: Praksis, Læring og Utvikling [Assessment Competence in School: Practice, Learning and Development]*, Oslo: Universitetsforlaget, 42–58.

- Smith, K (2016) Cooperative Learning about Assessment for Learning, in Laveault, D and Allal, L (Eds) *Assessment for Learning: Meeting the Challenge of Implementation*, Cham: Springer, 181–197.
- Smith, K and Engelsens, K S (2012) Developing an Assessment for Learning Culture in School: The Voice of the Principals, *International Journal of Leadership in Education* 16 (1), 106–125.
- Sylte, A L (2014) Vurdering for yrkesrelevant oppl ring [Assessment for vocational training], *Nordic Journal of Vocational Education and Training* 4, 1–18.
- Taylor, L (2013) Communicating the Theory, Practice and Principles of Language Testing to Test Stakeholders: Some Reflections, *Language Testing* 30 (3), 403–412.
- Tsagari, D (2021) Gauging the Assessment Literacy Levels of English Language Teachers in Norway, *European Journal of Applied Linguistics and TEFL* 10 (1), 161–191.
- Tveit, S (2014) Educational Assessment in Norway, *Assessment in Education: Principles, Policy & Practice* 21 (2), 221–237.
- Vogt, K and Tsagari, D (2014) Assessment Literacy of Foreign Language Teachers: Findings of a European Study, *Language Assessment Quarterly* 11 (4), 374–402.
- Wenger, E (1998) *Communities of Practice: Learning, Meaning, and Identity*, Cambridge: Cambridge University Press.
- Wiliam, D (2018) *Embedded Formative Assessment* (Second edition), Bloomington: Solution Tree Press.

14

'A course of our own': Reflections following a language assessment literacy (LAL) course

Ofra Inbar-Lourie

Tel-Aviv University, Israel

This reflective piece draws on experiences obtained while teaching a language assessment literacy (LAL) course to pre-service teachers of different languages as part of a teacher certification program in a university in Israel. The insights gained highlight:

- The language assessment connection
- The benefits of conducting language assessment training in specific frameworks
- The role of critical language assessment and active relevant learning and research for the acquisition of LAL and language assessment identity
- The acquisition of LAL expertise as a situated paradigm from both language and assessment perspectives

מאמר רפלקטיבי זה מתבסס על התנסויות בהוראת אוריינות הערכה שפתית לפרחי הוראה במסגרת הכשרתם כמורי שפות וכחלק מלימודיהם לקראת תעודת הוראה באוניברסיטה בישראל. התובנות שהתקבלו מהתנסויות אלה מדגישות את ההיבטים הבאים:

- הקשר בין שפה להערכה
- היתרונות של קיום מסגרת נפרדת להכשרה להערכה שפתית
- תפקידה של למידה ביקורתית, פעילה, רלוונטית ומחקרית לרכישת אוריינות הערכה שפתית ופיתוח זהות של מעריכים שפתיים

רכישת מומחיות באוריינות הערכה שפתית כפרדיגמה הקשרית, הן מנקודת המבט השפתית והן מנקודת המבט ההערכתית.

Background

I enter the university classroom on a hot humid day at the very start of the spring semester 1998, to teach a course on language assessment. About 45 prospective language teachers are awaiting. The course forms part of their teacher certification program, delivered in Hebrew. I have taught the course numerous times before, mostly to English as a Foreign Language (EFL) teacher groups in English, and am a bit hesitant considering the mixed audience, not knowing which way this will go. I first introduce myself and briefly describe my personal journey as a language tester, and then ask the students to introduce themselves, and name their language specializations and background in assessment. The majority are studying to teach either Arabic as a Second Language (ASL) to Hebrew speakers or EFL. There are also prospective Hebrew and Arabic language arts teachers, and teachers of French as an additional language. Few have previous training or experience in conducting classroom assessment. They are all, however, experts in being assessed, having experienced classroom and external testing as students in the K-12 and academic systems.

I ask why they think all the language teachers are taking this course together rather than joining the regular assessment course for other subject teachers, and what they expect to gain out of the course in this format. The class is silent. One student raises her hand. 'I am studying to become a Lashon teacher', she says in obvious dissatisfaction, as if the rest is obvious. Lashon ('tongue' in Hebrew) is used (among other things) to refer to Hebrew language arts as a discipline in the educational system with a focus on language structures. My bewildered gaze warrants further explanations, which the speaker willingly and emphatically provides: 'Our students are Hebrew speakers and already know the language. We are not language teachers.' The teachers of L1 Arabic are also confounded, sharing a yet more restrained point of view. 'True', says one. 'But for us it's more complicated because the students use Spoken Arabic but we teach them the Standard Literary form.'¹ The class is buzzing with comments. We all embark on a long discussion on what it means to be a language teacher of different languages, teaching different learner groups for different purposes under different circumstances.

We talk about what language knowledge is, and what our common and unique goals for teaching the languages are. The students agree on some similarities. For one, they are worried to varying degrees about evolving dynamic language changes, particularly the impact of the emergence of digitalization (at the time), and the implications for language accuracy,

¹ Arabic is a diglossic language consisting of Modern Standard Arabic common to all speakers of Arabic and spoken Arabic which differs geographically (Ferguson 1959).

standardization and teaching. Inevitably, themes of language in society and politics in the multilingual Israeli context emerge, especially with regard to the status of the languages represented in our class. The conversation highlights the asymmetry between Hebrew and Arabic in Israel in general and particularly in the educational system (Or and Shohamy 2016), the dominant role of English and the declining status of French (Spolsky and Shohamy 1999), which competes with Arabic as the students’ choice of an additional language (Inbar-Lourie, Donitsa-Schmidt and Shohamy 2000). Other issues concern the maintenance of heritage languages including those of some of the participants (e.g., Russian), and the difficulties of teaching a diglossic language. The 90 minutes of class time go by fast, and the word ‘assessment’ is hardly mentioned, except with reference to external large-scale tests that dominate the school ecology (Inbar-Lourie and Shohamy 2021). We part with a thinking task: Why a course on *language* assessment?

The topic unfolded over the next weeks as we began to jointly disentangle and unwrap the assessment issues along with discovering language constructs, working on assessing the different language skills and genres. We gradually established common ground and discourse, combining language and assessment matters. The students read and came back with questions, reviewed assessment instruments and analyzed their items in groups. The theoretical underpinnings often emerged inductively from the practice or vice versa. I observed the homogeneous or mixed language groups working towards gaining a language assessment knowledge base. Yet, many of the concepts remained abstract: achievement and proficiency, reliability and validity, formative and summative assessments. Moreover, that magic ingredient which creates excitement and enthusiasm as well as commitment to language assessment was still missing. Reminiscences of my days as a student in the assessment course (or ‘language testing’ course as it was called), conjured up memories as to where and how my passion for the field was kindled. The recollections include the wide spectrum through which tests were presented as socially and politically situated tools, rather than merely measurement instruments. There are also memories of participation in research, of constructing and trying out assessment tools in class, and of heated discussions and lobbying against test misuse. All of these enacted feelings of belonging to the language assessment community and planted the seeds for the formation of a language tester identity. As in all learning events, relevant experience is the name of the game. Authentic. Real.

The approach

Phase One: Relevance

The class was encouraged to take on an active, personalized and involved role. We started by giving room to the students' previous and current experiences as test-takers and providers, and analysed the consequences of these experiences. Students shared media coverage of the external exams so prevalent in the Israeli educational system. They developed critical awareness of the washback of tests including the time spent on exam preparation, curricular considerations, as well as test bias and the difficulties students encounter when labeled based on exam results. They debated the problematic psychometric qualities of the Israeli matriculation exams (that determine entry to academic studies), for the tests are not piloted prior to administration, jeopardizing test quality and standardization. Simultaneously, they acquired tools for determining the quality of language tests and their scoring, using procedures such as item difficulty, item discrimination and inter and intra-rater reliability. They also learned how to interpret results and communicate them to their students, colleagues and school administrators to point out levels of heterogeneity in their classrooms, revise teaching plans, request additional resources based on assessment results and provide explanations. They then tried out all of these tools and newly gained skills in real time with their practicum teachers, and found that more often than not they had become the source of expertise in this respect.

We worked on finding alternative diverse means for assessment focusing on assessing speaking and listening comprehension, neglected skills seldom assessed in the Israeli language classroom and included only in some of the external language tests. They learned from each other with different languages leading the way for particular scenarios, developing tools and assessment criteria. One of the leading language groups was ASL, which is quite unusual as it is more common for EFL to provide models for language teaching and assessment, with communicative goals and a tradition of performance tasks. This shift can be traced to a pilot research study on assessing spoken Arabic, in which the Arabic pre-service teacher group took an active role as is described below.

Phase Two: Participatory research and hands-on experimentation

The pilot research was part of a study conducted on teaching spoken Arabic, in an attempt to encourage teachers to engage in teaching the spoken language as a communicative tool in Jewish primary schools (Donitsa-Schmidt, Inbar-Lourie and Shohamy 2004). Though about 40% of Hebrew-speaking

students study Arabic (especially in Junior High Schools), the level of proficiency achieved is very low. It is claimed that one of the reasons for this situation is the policy decision to teach Modern Standard Arabic (MSA) based on its uniformity, versus the spoken dialects, which allow for localized communication (Amara, Azziza, Mor-Sommerfeld and Hertz-Lazarowitz 2008, Or and Shohamy 2016). In an effort to make the language more meaningful and communicative, a new program for teaching the Palestinian Arabic dialect was initiated by the Tel-Aviv Jaffa municipality. The program targeted young learners in an attempt to improve their motivation towards studying Arabic, as well as avoid stereotypical perceptions of its speakers in view of the complex political and fragmented reality in the Israeli society (Donitsa-Schmidt et al 2004).

I therefore teamed up with the Arabic pedagogical instructors who emphasized in their classes methods for teaching the spoken language. Under our joint guidance, our prospective Arabic teachers devised speaking tasks intended to elicit and assess the Arabic oral abilities of primary school pupils in Grades 4 to 6 (10 to 12 years old) participating in the program. Following a literature review on assessing the oral skills of young language learners and a fruitful discussion, we reached two decisions: (a) the elicitation tasks would vary to reflect the teaching and allow for different language output, and would include an interview, a picture description and a role-play activity; (b) all three tasks would be assessed according to a holistic rating scale with relevant criteria. The design of the assessment scheme and procedures took into consideration the young learners’ age and limited language knowledge. We all collaborated on a pilot version and went about constructing the tasks and a six-level scale. Both were piloted hands-on as part of the data collection in the schools. The process generated authentic assessment dilemmas and challenges which we needed to face and solve, e.g., decisions concerning the holistic scoring with possible overlaps; the test-takers’ idiosyncratic features including their ability to cope with the tensions of spoken assessments and interactions with the assessors; use of first language; differential performance on the three tasks, and more.

The pre-service teachers of Arabic shared the entire assessment experience with the assessment class, both ongoing for relevant feedback and at the final stages, allowing for group and personal reflective analysis. This sparked vivid interest even among the most skeptical pre-service language teachers, who followed up by reflecting on similar hands-on endeavors. We set out on small classroom-based research projects, whereby the prospective teachers created performance tasks within each language, exploring how the accuracy and communicative aspects interact, and how to integrate teaching with assessment in a meaningful manner. The diverse performance tasks explored different genres and registers, designed to follow process as well as product skills, and to represent both formal and informal text types and registers.

Examples include following and providing instructions (from assembling a kite to baking cookies, delivering a speech and creating a portfolio), listening to the news, weather forecasts, and talk shows and orally producing the same genres, and conducting surveys (e.g., on students' assessment preferences).

The prospective teachers and now assessors raised questions throughout the process, especially regarding the teaching–learning–assessment cycle, i.e., how to provide constructive feedback to improve learning but also inform teaching. Mutual learning occurred across languages. Obscure concepts such as validity and reliability of closed and open item formats came to life, united under common understandings. Furthermore, initial recognition of the students' future dual roles as both language teachers and language assessors filtered in, adding another dimension to the development of their professional identity. The initial doubts as to the need for a special assessment course for language teachers (some still probing their professional identity), had given way to the establishment of a community of practice which, in the words of some of the participants, appreciates having 'a course of our own'.

Insights: Current understandings of LAL

The course described took place close to 25 years ago. Many courses have followed intended for diverse audiences, both general and language-focused, aligned with current thinking in the field. The initial insights gained have reoccurred repeatedly and consistently, and have assisted in formulating my basic thoughts on LAL and its attributes as well as on the features of language assessment courses (Inbar-Lourie 2008). Though the LAL concept has since evolved and expanded through abundant studies by researchers in different contexts and through deliberations as to its essence (Inbar-Lourie 2017), the basic understandings from the course described still resonate in my LAL conceptualizations, especially with regard to the apprenticeship of language testing experts.

The recollections outlined above, as well as those from other course formats, have reinforced my belief that the language assessment synergy is essential in the construction of language assessment expertise. This leads to my first insight and strong conviction that the mentoring track for LAL should include a course specifically designed for language assessors, one that can address language-related relevant themes. A noticeable current example that comes to mind is the translanguaging motif as a conceptual paradigm (García and Wei 2014), and its implications to translanguaging pedagogy (Cenoz and Gorter 2022), and hence to conducting language assessment. Only an assessment course with a language orientation can consider these implications to assessment theory and practice. The same holds true for other issues, like current conceptualizations of the native speaker construct.

This language and assessment fusion holds true for language teachers and experts, but also for others involved in situations where the language plays a major role in the teaching, learning and assessment process, such as in the case of content lecturers in English as a Medium of Instruction (EMI) programs whose assessment literacy will include awareness of the role language plays in accessing academic content (Inbar-Lourie 2022).

The second insight refers to the importance of exposing the mentored language assessment experts to the powerful scholarship of critical language testing. Viewing assessment in general and language assessment in particular through this lens broadens the assessment perspective to include vital existential issues of social justice and commitment. Our field is indebted to a number of scholars who have relentlessly advocated for this direction. I was especially influenced by my mentor, Professor Elana Shohamy, who is continually advancing theory and research in critical language assessment following the publication of her seminal book *The Power of Tests* (2001), accentuating the need for equitable fair assessment practices.

The third element in planning language assessment courses is directly connected to the previous two, and that is the activist experiential ‘learning by doing’ component (Kleinsasser 2005). Introducing prospective language teachers to language assessment is a lengthy process. Becoming an involved language assessment practitioner, and eventually taking on a language assessor’s identity, is an even more challenging undertaking that necessitates not only acquisition of the relevant concepts, terms and procedures, but also the ability to process and actively utilize them within macro and micro contexts. Hence the emphasis on actually practicing classroom assessment as part of the learning experience. Additionally, taking part in any form of research that investigates the role of language assessment and provides experience in facing and solving challenging dilemmas is strongly recommended as part of the development of LAL expertise.

Last but not least: LAL is a situated phenomenon, from both the sociolinguistic and assessment perspectives. This was evident in the course interactions described here, as the status of the various languages in the Israeli context as well as the heavily accentuated testing agendas both sketched the participants’ perceived positioning and assessment needs even prior to their undertaking official teaching responsibilities. Thus, though language assessment courses share common LAL content topics, the local language and assessment ecologies and cultures need to be contemplated as part of establishing the dynamic and evolving foundation of the language assessment literacies. On this note, I would like to express my appreciation and gratitude to my students and colleagues for generously sharing with me their opinions, experiences and wisdom in the various language assessment courses throughout the years.

References

- Amara, M, Azziza, F, Mor-Sommerfeld, O and Hertz-Lazarowitz, R (2008) *Teaching Arabic as a Foreign Language in Israel's Jewish Schools: Challenges and Recommendations*, Haifa University, The Jewish-Arab Center: The Program for Bilingual Education.
- Cenoz, J and Gorter, D (2022) *Pedagogical Translanguaging*, Elements in Language Teaching, Cambridge: Cambridge University Press.
- Donitsa-Schmidt, S, Inbar-Lourie, O and Shohamy, E (2004) The effects of teaching spoken Arabic on students' attitudes and motivation in Israel, *The Modern Language Journal* 88, 217–228.
- Ferguson, C A (1959) Diglossia, *Word* 15 (2), 325–340.
- García, O and Wei, L (2014) *Translanguaging: Language, Bilingualism and Education*, Basingstoke: Palgrave Macmillan Pivot.
- 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 (2017) Language assessment literacy, in May, S (Ed) *Volume on Language Assessment*, Encyclopedia of Language and Education, Cham: Springer Science + Business Media LLC, 257–270.
- Inbar-Lourie, O (2022) EMI programs and formative assessment: Implications for the assessment literacy of content lecturers, *The Journal of EMI* 2, 204–231.
- Inbar-Lourie, O and Shohamy, E (2021) The conflict and consequences of two assessment measures in Israel: Global PISA vs. the National MEITZAV, in Lantaigne, B (Ed) *Challenges in Language Testing around the World*, Cham: Springer, 191–200.
- Inbar-Lourie, O, Donitsa-Schmidt, S and Shohamy, E (2000) Students' motivation as a function of language learning: The teaching of Arabic in Israel, in Dörnyei, Z and Schmidt, R (Eds) *Motivation and Second Language Acquisition*, Manoa: University of Hawaii Press, 289–303.
- Kleinsasser, R C (2005) Transforming a postgraduate level assessment course: A second language teacher educator's narrative, *Prospect* 20, 77–102.
- Or, I G and Shohamy, E (2016) Asymmetries and inequalities in the teaching of Arabic and Hebrew in the Israeli educational system, *Journal of Language and Politics* 15 (1), 25–44.
- Shohamy, E (2001) *The Power of Tests: A Critical Perspective on the Uses of Language Tests*, London: Longman.
- Spolsky, B and Shohamy, E (1999) *The Languages of Israel: Policy Ideology and Practice*, Clevedon: Multilingual Matters.

15

Advancing professionalisation: The achievement of language assessment literacy

Albert Weideman

University of the Free State, South Africa

This reflection:

- Traces the growth of interest in language assessment literacy (LAL) over the past two decades
- Considers how LAL has come to be conceptualised and characterised over that time
- Views language tests from the viewpoint of technical design as an angle from which to evaluate the historical significance of LAL
- Draws upon a set of systematic criteria to analyse these chronological developments, taken from a philosophy of technology which has proved to be theoretically productive for applied linguistics
- Suggests that LAL has brought about several advances in the professional understanding of language testing

Hierdie terugskouing:

- ondersoek die groeiende belangstelling in taalassesseringsgeletterdheid (TAG) oor die afgelope twee dekades
- verken hoe TAG gekonseptualiseer en gedefinieer is gedurende daardie tydperk
- beskou taaltoetsing vanuit die perspektief van tegniese ontwerp, as 'n invalshoek waaruit ons die historiese betekenis van TAG kan evalueer
- put uit 'n stel sistematiese kriteria, afkomstig uit 'n filosofie van tegniek wat geblyk het teoreties produktief te wees vir die toegepaste taalkunde, ten einde hierdie kronologiese ontwikkelings te analiseer
- bevind dat TAG op verskeie punte vooruitgang meegebring het in die professionele begrip van taaltoetsing

Growing interest in language assessment literacy

In evaluating the historical significance of language assessment literacy (LAL), we need more than a chronological description of the developments it stimulated. An issue grows into significance when it comes to be viewed as leading to a potential advance in a professional realm, such as the design of language tests. ‘Significance’ in this professional, historical sense means expressed or signified influence. Influence, in turn, is what drives history; influence, once exercised, means change; and desirable change, finally, indicates an advance, whereas undesirable change means regression. As we evaluate our professional history, we gauge the ups and the downs, and this reflection will argue that LAL constitutes a historical advance for the profession. It is an up rather than a down. Wider issues of language assessment are also relevant, but this reflection will focus on language test design.

Though there are references to some earlier work in the literature (e.g. Taylor 2009), there has been unabated attention to LAL since the second decade of the 21st century. Special volumes of language testing journals have been dedicated to it (see Taylor (2013) for a summary of one), and it features prominently in reference works (e.g. Inbar-Lourie 2017). It is a conspicuous theme at professional gatherings of language testing specialists, such as the annual conference of the International Language Testing Association (ILTA) the Language Testing Research Colloquium (LTRC). At LTRC 2017, for example, the main conference theme was ‘Language assessment literacy across stakeholder boundaries’, and at LTRC 2018 several papers took that discussion further.

Taylor (2013) ties rising interest in LAL to a historical trend in the late 20th century to consider all manner of ‘literacies’, be they media, technological, cultural, computer, or financial, in order to describe a competence in the field denoted. This observation signifies that we are never immune to historical trends. Chronological descriptions of the development of LAL itself note that its initial concern was with language teachers’ knowledge of testing principles and professional practices (e.g. Fulcher 2012, Weideman 2019a; also Inbar Lourie 2017:259, 262f.). From the start, however, though that was the primary focus, teachers’ competence was not its only concern. Taylor (2009) observes that it applies equally to the knowledge of assessment and practices of users of test scores in a number of institutional environments. Those institutions mostly are the state and universities, the typical seats of high-stakes decisions. For the purposes of this contribution, the eventual emphasis will fall on how LAL considerations may affect schools and universities in South Africa.

Conceptualising and characterising LAL

Baker (2016) identifies the familiar components of the construct of LAL. These are conceptual knowledge of language assessment and its procedural practices, with a policy or ethical dimension, or both, being added, following Taylor's (2013:410) dynamic or differential conceptualisation which includes various stakeholders (e.g. teachers, test developers, and administrative users of test results). Conceptual knowledge is explained as theory, principles or knowledge of language testing, while procedural practices include experience in the administration of tests.

What is of importance for this analysis, however, is that both Taylor (2013) and Baker (2016) identify a reciprocity between language testers and other stakeholders. For LAL to develop across a range of stakeholders, one would need 'a collaborative ongoing dialogue ... between language testing experts and non-specialist test stakeholders' (Taylor 2013:411). Thus, from an initial concern with kindred professionals (language teachers), the issue has now been extended to include a range of non-expert (in that respect) but still highly involved decision makers. Yet even with teachers, the reciprocity and collaborative dimension is clear: the report by Baker and Riches (2018:574) of their involvement with teachers concludes that both groups (teachers and testing professionals) benefitted from their collaboration. So the question is: How can such reciprocity and extension be further characterised, and what is their significance?

There is little doubt that shifting historical emphases within language assessment have contributed much to the extension of the concerns of language testers to include, for example, official decision makers. It is telling that the notions of 'transparency', 'accessibility', 'communication', 'accountability' and 'impact' figure a little more prominently in later reviews (Inbar-Lourie 2017, Taylor 2013) than concepts like reliability, validity, and construct. Our concerns have shifted towards cultural, social, political and ethical dimensions of language testing (McNamara and Roever 2006, Weideman 2017a). I shall argue that, in order to evaluate their significance, we need a theoretical framework to account for such shifts. Although the sociocultural turn in applied linguistics can be used to account for some of the shifts, another, complementary perspective can be found in a different theoretical framework originating in the philosophy of technology and technical development, to which I now turn.

Drawing on insights from the philosophy of technology: A theory of disclosure

Understanding LAL as an outcome of recent scholarship allows us to develop a more sophisticated understanding of its effect on language test design, and

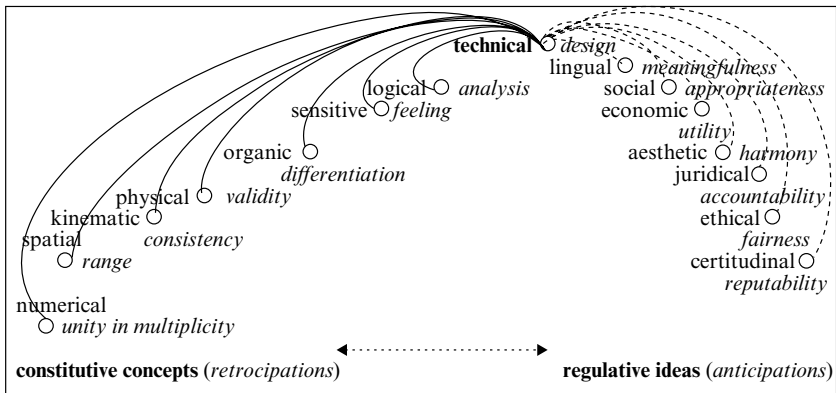
to consider whether it has contributed to a disclosure of design. ‘Disclosure’ as used here has a conceptual sense of ‘opening up’ or ‘deepening the meaning of’. This phenomenon can be characterised according to current perspectives in the philosophy of technology (Schuurman 2009, 2022), which argue that a key feature of the opening up and disclosure of a new development in applied linguistic design is its anticipation of the future. Schuurman (2022:83) calls this ‘conscious planning’, adding that ‘attention should be given to proactive reflection-ahead-of-time, instead of the current tendency to reflect after the event’. This echoes exactly what Taylor (2013:407) has observed: ‘Ideally the promotion and development of assessment literacy will be achieved “by design” rather than being a corrective afterthought.’ In other words, a well-developed understanding of the principles and practice of assessment needs to inform the process of test design from the earliest design phase right through to the delivery and implementation stages, proactively anticipating the likely future impact of the instrument as thoroughly as possible (Rambiritch 2012).

I have chosen Schuurman’s (2022:83) ‘cohesive, ethical evaluating framework’ since his work in the philosophy of technology supports a view of applied linguistics that can be productively employed to assess advances or to register regression in our designs of language tests. To view these planned measurements from the angle of design (Weideman 2017b) is the beginning of a constructive theory. The starting point is to abstract the technical mode of design that characterises applied linguistic interventions, and to examine the connections that this modality has with other dimensions of experience (Weideman 2023).

First, the theory therefore begins by acknowledging that our experience as applied linguists is structured to recognise a multiplicity of aspects or modalities (features/characteristics/modes): for instance, numerical, spatial, kinematic, physical, organic, sensitive, logical, technical, lingual, social, economic, aesthetic, juridical, ethical and certitudinal modes. Second, let us accept that these aspects may also serve as modes of explanation and, therefore, of conceptualisation. We can theoretically explain an object, state, event or phenomenon, for example a language test, from many modal angles: emotional, economic, social, juridical or ethical. When we view language tests from the angle of design, as is proposed here, one of these aspects, in this case the technical, stands out as our angle of approach, since, as applied linguists, our interventions are characterised by designing, shaping, forming, arranging, or planning of this particular kind of solution to a language problem. Third, this angle (in our case the technical) does not exist on its own, but is related to all the others mentioned in the list above. They are analogically reflected in it. Thus, the methodology for the application of the theory is: conceptually engage with these analogical reflections of other aspects within the technical in order to lay bare the basic concepts of applied linguistics as a discipline characterised by design.

Figure 1, adapted from Weideman (2009), diagrammatically presents the various modalities (numerical, spatial, kinematic, through to juridical, ethical and certitudinal) with which the technical mode of design is associated analogically. The notion of constitutive concepts and regulative ideas is explained below; the terms ‘retrocipations’ and ‘anticipations’ refer, respectively, to constitutive and regulative analogies of other modalities

Figure 1 Coherence of the technical modality with others (and their analogies)



within the technical. The regulative applied linguistic ideas, deriving from ‘anticipations’ of other spheres in our technical design activity, are the leading technical ideas guiding the meaningful disclosure of language test design, and therefore the focus of this analysis.

Fourth, these basic concepts allow the theorist to generate design principles. In Schuurman’s ‘integral framework of principles’ these connections among different modalities constitute a ‘helpful grid for ... responsible technical development’ (2022:81). The principles include the norm of effectiveness, as well as

the norms of *harmony* between continuity and discontinuity, between large-scale and small-scale, between integration and differentiation, ... the norms of clear information and open communication, ... the norms of stewardship, of efficiency, of doing universal justice to the “players” involved ... the norms of care and respect for ... everyone involved ... and the norms of service, trust, and faith. (Schuurman 2022:81; emphasis in original)

In order to understand how these different principles fit into a framework (presented diagrammatically in Figure 1), one needs to refer to an earlier analysis by Schuurman (2009:416ff.). Here, disclosure is achieved

when the technical mode, the nuclear meaning of which lies in design (*'het zwaartepunt'*, meaning its 'centre of gravity'), first anticipates the lingual modality, by expressing the design in the form of a blueprint and specifications. In test design, these blueprints come in the shape of test construct and specifications. Similarly, the technical anticipates the social mode in the teamwork that prepares for the implementation of the planned measurement, and in the appropriateness or 'fit' of the test for its intended population. The design principles that derive from such echoes of other modalities within the structure of the technical apply to the three prime language interventions, language course and policy development, as well as language assessments (Weideman 2017b:219–226). All adhere to general principles, converting them into design prescriptions that act as conditions for the plans and arrangements foreseen. The focus here, though, is on language test design.

The different design principles identified above by Schuurman (2022) can be related, respectively, to the way that the aesthetic dimension is anticipated by the technical (in fulfilment of the 'norm of harmony'); to how the lingual and social are further disclosed by the 'the norms of clear information and open communication'; to how the norms of stewardship and efficiency can open up the design with reference to the economic mode; how doing justice illustrates the anticipation of the juridical in our designs; how the principles of care and respect anticipate the ethical dimensions that our technical designs will have for users; and finally how, when 'the norms of service, trust and faith' are brought into play, we see an anticipation in our designs of the aspect of commitment and certitude.

Since they derive from technically stamped ideas, these principles serve as lodestars for our design of language interventions (Weideman 2009, 2019d, 2023). The regulative technical ideas from which they stem (Weideman 2009) act as guides to disclose the meaning of our designs. The principles thus derived fall into both regulative and constitutive categories: in the excerpt above, Schuurman mentions the norms of integration (a numerical echo in the technical) and differentiation (referring to the mode of organic life), of technical range (small-scale and large-scale), which reflects the spatial mode, and of effectiveness, a reference to the physical mode of energy effect. These, as is set out in Figure 1, are constitutive ('necessary') or founding principles for our designs. In language testing we are familiar with the analogically physical notion of technical validity, and the kinematic analogy of reliability, both constitutive building blocks on which our designs are necessarily founded (Van Dyk 2010, Weideman 2009, 2019c). However, when we evaluate the disclosure of the meaning of our designs, we should, in Schuurman's theory, look not only at constitutive building blocks, but at the regulative technical principles which, once they are given positive form in our designs of language interventions, indicate the degree to which they contribute to an unfolding and disclosure of the meaning of design.

Evaluating the contribution of assessment literacy

How can we, in light of the theory set out above, evaluate the historical contribution of LAL? Using Schuurman's framework, we may note several such disclosures of the technical function of design in language tests, particularly relating to ethical, economic and other regulative conditions.

Ethical gains: Consequences matter

Taking the ethical connection with the leading technical mode of design first, we may note in discussions of LAL the prominence of the technical impact of language tests (Taylor 2013:406f.). Though such impact relates primarily to the connection of the technical design with ethical considerations – 'impact' then intended as 'ethical consequences' – language test providers are rightly concerned also about the eventual social, economic and political effects of tests. Anticipating the consequences of their measurement results has opened up the meaning of test design to foreseeing ethical issues (Schildt, Deygers and Weideman 2023). We design with care and compassion, mindful of whether our tests have beneficial consequences for those who take them.

Economic gains: Usefulness enhanced

Besides these ethical anticipations of language test design, we also find in the literature references to economic considerations. Baker (2016:68) specifically mentions usefulness (by end users) as a criterion for involvement with university admission officials. Perhaps politicians may wish to employ language tests not as 'power' plays to keep immigrants out, but out of genuine concern to ensure the productive participation of new entrants in the economy and in civic life.

Juridical gains: Justice is more prominent

Such concerns with 'social justice' (Rawls 1990:16) remind us of how juridical considerations also impact language test design. This constitutes a further disclosure of the meaning of design. Rawls' (1990:15) observations on 'principles of justice' assign to language test designers the task of restoring and correction, of righting wrongs. In initially drawing attention to an undesirable gap in expertise among language testers and teachers, those who brought LAL to prominence strove to correct a rectifiable technical imbalance.

Lingual and social gains in transparency, accessibility and accountability

Rambiritch (2012) presents an analysis of further disclosures of the technical in examining whether informativity and transparency can be built into the design of a test. Transparency is here being re-articulated as the knowledge of language test-takers about the measurement. These are clearly echoes of the lingual aspect within our technical design domain. On such concerns, as Taylor (2013:406) notes, are built further unwrappings: the technically achieved greater accessibility and enhanced public communication about what a test measures. All of these advances, of bringing lingual and social considerations into reckoning in designing language tests, spring from our focus on LAL. And on them, in turn, relies our public accountability for the tests we design (Davies 2008, Weideman 2006, 2017a), another reflection of the juridical.

Certitudinal gains: Humility is indicated

Finally, the certitudinal echoes within test design are fronted in the notion of LAL. Baker (2016:82) is rightly critical of imposing a deficiency model on those professionals in other fields with whom language test designers interact. We cannot believe that we hold the ultimate wisdom if we wish to learn from others (as we should). In short, language testers need to approach those interactions with appropriate humility, and an acknowledgment of their limitations.

Nonetheless, when we apply Schuurman's (2009, 2022) theory of technical disclosure we can identify lingual, social, economic, juridical, ethical and certitudinal anticipations that deepen (in the sense of disclosing and enhancing) our designs. These sensitivities relate either directly or obliquely to LAL. For this reason language testers can be rightly proud of taking a step up in the professionalisation of their field. One must remember, too, that the interactions we have recorded spring from language test designers, more than from those we interact with. Humility is in order, but so is some gratification in our anticipation and intention of designing responsibly.

Professionalisation and further challenges

The consideration of LAL has opened up new professional avenues for language test designers. We have managed to link language testing with another prime applied linguistic artefact, language policy (Lo Bianco 2001, 2014). Now we need to acknowledge, when we take a broader view, that we have not yet reached the goal of responsible design of all language interventions. In South Africa, for example, we still have substantial work

to do. I conclude this discussion with some prospects of what needs to be undertaken, that will clearly emphasise how interwoven LAL is with both language policy and language course design.

First, with due regard for South Africa's multilingual context, and in order to register the kind of professional strengthening discussed above, we need to undertake research on assessment policy literacy and assessment literacy at secondary school level. This is needed as rectification of the serious misalignment between school language curricula and exit examinations (Weideman 2019b, Weideman, Du Plessis and Steyn 2017). Seeking such technical harmony relates once more to the theoretical framework utilised above. In this case, we seek to give positive shape to a design principle emanating from the connection between the technical mode of design and the aesthetic, to achieve the planned alignment of these different applied linguistic designs (Weideman 2023).

Second, one may consider the lingual, juridical and ethical consequences of test design. High-stakes tests in South Africa are concerned with access to university (Deygers and Malone 2019, Deygers, Van den Branden and Van Gorp 2017). Two sets of research questions need to be answered if we are to gain an understanding of how university admission policies and practices employ tests of language ability:

- (a) Do the outcomes of language teaching at school indicate sufficient readiness for tertiary education? How does the use of measures of academic literacy affect access to tertiary education? How do admissions officials in fact interpret results (e.g., mechanically or with discretion)?
- (b) What further opportunities for language development are offered after assessment? What are the further challenges of aligning institutional policy at university with the realities of the academic literacy levels of prospective students? What is the level of language policy literacy?

This research needs to be undertaken to first gain an understanding of both practices and policies. Though test practices and designs have been recorded and in some instances quite comprehensively analysed (see the bibliography of Network of Expertise in Language Assessment (NExLA) 2022, Weideman, Read and Du Plessis (Eds) 2021), gaps remain. A better understanding should prepare us to give professional advice that is backed up by research. In short: we have a context here that cries out for the kind of professionalisation that this contribution has identified and celebrated.

Acknowledgement

I am grateful especially to the editors, who have made helpful suggestions. The final version remains my responsibility.

References

- Baker, B (2016) Language assessment literacy as professional competence: the case of Canadian admissions decision makers, *The Canadian Journal of Applied Linguistics* 19 (1), 63–83.
- Baker, B and Riches, C (2018) The development of EFL examinations in Haiti: collaboration and language assessment literacy development, *Language Testing* 35 (4), 557–581.
- Davies, A (2008) Accountability and standards, in Spolsky, B and Hult, F M (Eds) *The Handbook of Educational Linguistics*, Oxford: Blackwell, 483–494.
- Deygers, B and Malone, M E (2019) Language assessment literacy in university admission policies, or the dialogue that isn't, *Language Testing* 36 (3), 347–368.
- Deygers, B, Van den Branden, K and Van Gorp, K (2017) University entrance language tests: a matter of justice, *Language Testing* 35 (4), 449–476.
- Fulcher, G (2012) Assessment literacy for the language classroom, *Language Assessment Quarterly* 9 (2), 113–132.
- Inbar-Lourie, O (2017) Language assessment literacy, in May, S (Ed) *Volume on Language Assessment*, Encyclopedia of Language and Education, Cham: Springer Science + Business Media LLC, 257–270.
- Lo Bianco, J (2001) Policy literacy, *Language and Education* 15 (2–3), 212–227.
- Lo Bianco, J (2014) Dialogue between ELF and the field of language policy and planning, *Journal of English as Lingua Franca* 3, 197–213.
- McNamara, T and Roever, C (2006) *Language Testing: The Social Dimension*, Oxford: Blackwell.
- Network of Expertise in Language Assessment (NExLA) (2022) *Bibliography*, available online: nexla.org.za/research-on-language-assessment/
- Rambiritch, A (2012) *Transparency, Accessibility and Accountability as Regulative Conditions for a Postgraduate Test of Academic Literacy*, PhD thesis, University of the Free State.
- Rawls, J (1990) *A Theory of Justice* (Revised edition), Cambridge: The Belknap Press of Harvard University Press.
- Schildt, L, Deygers, B and Weideman, A (2023) Language testers and their place in the policy web. Under review for *Language Testing*.
- Schuurman, E (2009) *Technology and the Future: A Philosophical Challenge*, Grand Rapids: Paideia Press, translated by H D Morton. [Originally published in 1972 as: *Techniek en Toekomst: Confrontatie met Wijsgerige Beschouwingen*, Assen: Van Gorcum].
- Schuurman, E (2022) *Transformation of the Technological Society*, Iowa: Dordt Press.
- Taylor, L (2009) Developing assessment literacy, *Annual Review of Applied Linguistics* 29, 21–36.
- Taylor, L (2013) Communicating the theory, practice and principles of language testing to test stakeholders: some reflections, *Language Testing* 30 (3), 403–412.

- Van Dyk, T (2010) *Konstitutiewe Voorwaardes vir die Ontwerp en Ontwikkeling van 'n Toets vir Akademiese Geletterdheid*, PhD thesis, University of the Free State.
- Weideman, A (2006) Transparency and accountability in applied linguistics, *Southern African Linguistics and Applied Language Studies* 24 (1), 71–86.
- Weideman, A (2009) Constitutive and regulative conditions for the assessment of academic literacy, *Southern African Linguistics and Applied Language Studies Special issue: Assessing and Developing Academic Literacy* 27 (3), 235–251.
- Weideman, A (2017a) Does responsibility encompass ethicality and accountability in language assessment?, Special edition of *Language & Communication* 57, 5–13.
- Weideman, A (2017b) *Responsible Design in Applied Linguistics: Theory and Practice*, Cham: Springer.
- Weideman, A (2019a) Assessment literacy and the good language teacher: four principles and their applications, *Journal for Language Teaching* 53 (1), 103–121.
- Weideman, A (2019b) Definition and design: aligning language interventions in education, *Stellenbosch Papers in Linguistics Plus* 56, 33–48.
- Weideman, A (2019c) Degrees of adequacy: the disclosure of levels of validity in language assessment, *Koers: Bulletin for Christian Scholarship* 84 (1), 1–15.
- Weideman, A (2019d) Validation and the further disclosures of language test design, *Koers: Bulletin for Christian Scholarship* 84 (1), 1–10.
- Weideman, A (2023) The practicality of theory: assessment and applied linguistics, under review for *SPIL Plus*.
- Weideman, A, Du Plessis, C and Steyn, S (2017) Diversity, variation and fairness: equivalence in national level language assessments, *Literator* 38 (1), available online: literator.org.za/index.php/literator/article/view/1319/2293
- Weideman, A, Read, J and Du Plessis, T (Eds) (2021) *Assessing Academic Literacy in a Multilingual Society: Transition and Transformation*, New Perspectives on Language and Education 84, Bristol: Multilingual Matters.

16

Assessment literacy theory: Pragmatics, complexity and transdisciplinarity

Christopher DeLuca
Queen's University, Canada

In this chapter, I describe core tenets of assessment literacy. My primary arguments are:

- Recent trends to shift assessment theory, practice, and literacy from a general discourse to a disciplinary one may result in unnecessary balkanization and atomization of assessment thinking, missing the spirit of assessment work
- Assessment literacy theory should, by necessity, remain discipline agnostic
- Three tenets of assessment literacy theory are: (a) assessment literacy theory is inherently pragmatic, (b) assessment literacy is a transdisciplinary theory, and (c) assessment literacy theory is inevitably complex

Assessment literacy theory: Pragmatics, complexity, and transdisciplinarity

A few years ago, at the American Educational Research Association Conference in Toronto, I attended a presentation session focussed on assessment in the disciplines. The presenters were calling to localize assessment theory within subject disciplines (specifically mathematics, science, and language), arguing that assessment theory and practice shifted from one disciplinary context to the next, akin to pedagogical content knowledge. Similar calls have been made in research studies and are particularly prominent in the language learning community, with efforts to isolate theories specific to language assessment and assessment literacy (see for example Lan and Fan 2019). I have become increasingly fascinated by these efforts and how they might balkanize the field of assessment writ large, leading to the siloing of assessment and causing educators to focus inwards on

their own specific domain instead of maintaining a broader interdisciplinary perspective. In this reflection, I share my cautionary thoughts on this trend, and offer instead a view that values assessment as inherently and necessarily transdisciplinary, pragmatic, and complex.

I begin my reflection by considering the persistent calls to translate assessment for learning (AFL) theory into domain-specific disciplinary theory. These calls stem, in part, from Bennett's (2011:15) critical observation as to 'whether formative assessment can be maximally effective if theory and development are focused at a domain-independent level'. He goes on to note that 'to be maximally effective, formative assessment requires the interaction of general principles, strategies, and techniques with reasonably deep cognitive-domain understanding'. However, despite Bennett's claims, and the work of several researchers who have since endeavoured to tailor AFL theory to disciplinary contexts, the theory remains, by and large, domain-independent and widely articulated through general principles. Why? Arguably, several factors maintain AFL theory as a cross-disciplinary enterprise.

Firstly, AFL research – and in fact, much of assessment and assessment literacy research – is at its core pragmatic in nature. By its form and purpose, assessment research is driven by pressures of the practical: an intense interest to support student learning through assessment *practice*. Hence, the general principles associated with AFL are inherently and epistemologically rooted in practice, and one might question whether it is the role of researchers or practitioners to translate AFL principles to specific contexts. Could maintaining AFL theory as a set of domain-independent axioms be more generative for educators by giving them freedom for enactment? Even more fundamentally, we might ask whether AFL principles should be specifically calibrated to disciplinary contexts, as initially suggested by Hodgen and Marshall (2005), particularly given the evolving nature of curricula and the vast number of curricular traditions and contexts that exist across jurisdictions.

In addition to being pragmatic, AFL is by design a transdisciplinary theory. As recognized by Ramadier (2004), transdisciplinarity involves unity in the production of knowledge, operating between, across, and beyond any discipline. It is intended to be flexible in its disciplinary (and cross-disciplinary) uses and applications. By operating as a set of general assessment principles for the improvement of learning, AFL has the capacity to move with curricula, reforms, and inventions, and across levels (i.e., grades). As curricula take flight in new directions, so too can AFL. For instance, consider the necessary potential practical adaptations to AFL within a math curriculum from Grade 1 to Grade 12, as learning progresses from simple to more complex mathematical literacy. Or consider how the practice of AFL might morph within integrated Science, Technology,

Engineering and Math (STEM) classrooms or play-based kindergarten contexts. Or the hundreds, if not thousands of curricular contexts across jurisdictions, levels, disciplines, and educational traditions that demand AFL – and indeed much of assessment theory – to be transdisciplinary. It is precisely the flexibility inherent in a transdisciplinary theory that enables teachers to interpret, adapt, and tailor AFL principles to their pedagogical approach and curricular context.

Finally, Marshall and Drummond (2006) famously made the distinction between enacting the letter versus the spirit of AFL. The *letter* of AFL represents a prescriptive implementation, with a one-to-one correspondence between theory and specified practices. In contrast, the *spirit* of AFL represents a more holistic adoption and absorption of assessment principles in one's pedagogy. I suspect that when AFL is understood as letter, there is greater desire to document its translation to specific disciplines; however, when AFL is taken up in spirit, such translation guides are less valuable as it is teachers' pedagogical practices that fundamentally shift toward a feedback-driven, student-centered approach. While discipline-dependent theory-to-practice translation guides might be valuable for some teachers as they learn about AFL and in their progression toward adopting the spirit of AFL, they can equally run the risk of imprinting the letter of AFL at the cost of its spirit. Moreover, adopting the spirit of AFL asks teachers to fundamentally re-examine how they teach, and to blend their personal pedagogy with AFL principles; such a view endorses individualized uptake of AFL across classrooms.

My thinking on AFL above is illustrative of my broader argument toward assessment literacy theory, which is the purpose of my reflective contribution in this book. My driving point is that assessment literacy, like AFL, is and likely ought to be a matter of general principle rather than a theory anchored to disciplinary knowledge and traditions. I take as my basis, as I often do, Willis, Adie and Klenowski's (2013:242) definition of assessment literacy:

Assessment literacy is a dynamic context-dependent social practice that involves teachers articulating and negotiating classroom and cultural knowledges with one another and with learners, in the initiation, development and practice of assessment to achieve the learning goals of students.

Inherent to this definition is the imperative for teachers to *negotiate* assessment knowledges in relation to their unique contexts. Theories of assessment literacy since that of Willis et al have further emphasized an interconnected orientation of praxiological, conceptual, and socioemotional dimensions that influence teachers' assessment work (Pastore and Andrade 2019). In other words, assessment literacy involves a teacher's understanding

and negotiation of assessment knowledge, its translation to practical contexts, and an attention to the social and emotional labor of assessment. Similarly, Looney, Cumming, van Der Kleij and Harris (2017) offer a view of assessment literacy rooted in teacher identity. In their view, assessment identity is comprised of a teacher's feelings, beliefs, knowledge, confidence, and perceived role toward assessment. While these dimensions are inevitably shaped by the disciplinary context in which the teacher works, discipline is not the sole nor driving factor shaping teachers' assessment literacy. Assessment literacy is far more than a monistic view of a teacher's knowledge of assessment standards and techniques in a specific discipline. In fact, one of the gross oversimplifications of the concept, in my view, is the reduction of assessment literacy to simply assessment skill.

To abstract further, assessment literacy theory maintains – arguably by necessity – several core features, which parallel those observed in our discussion of AFL, and which I will explicate further here. My aim is to enlarge the tenets of assessment literacy theory, to think beyond the specific factors, dimensions, and knowledges that comprise teachers' assessment work and which have been documented well elsewhere (see Xu and Brown 2016), and to consider the underpinning assumptions of the theory itself. Specifically, the three tenets that repeatedly surface in my consideration of assessment literacy theory are: (a) assessment literacy theory is inherently pragmatic, (b) assessment literacy is a transdisciplinary theory, and (c) assessment literacy theory is inevitably complex.

Inherently pragmatic

Like AFL, assessment literacy theory is ultimately preoccupied with assessment work in schools and the precursors, sociocultural conditions, and the contextual, relational, and personal influences on that work. For educators, assessment literacy is about the driving factors on teachers'¹ classroom assessment actions and the resulting experiences of assessment for students in schools. The essential qualities of pragmatism I wish to emphasize here are its (a) practical imperative, and (b) focus on experience (James 1907, Ormerod 2006). The practical imperative ties cognition, emotion, experience, and all other drivers of teachers' assessment approaches to their actions. It suggests that the work of assessment literacy theory is to understand the

¹ Note that my emphasis is on classroom teachers; however, the notion of assessment literacy can be applied, and has value, to other education stakeholders including students, parents, administrators, as well as those outside the formal system of education such as those responsible for citizenship policy and immigration, language assessment, professional certification, admissions, etc. Importantly, assessment literacy – as applied to any of these stakeholders – involves negotiating assessment knowledges with others and in relation to contexts.

dimensions, conditions, and contexts that shape teachers' practices. And from practice, comes experience, both for students and teachers.

Theories of experience in education are numerous, most notably the work of Dewey (1938/1959), which recognizes that students' experiences (i.e., what they see, hear, feel, and do) shape their emotions, learning, and their future actions and interactions, creating a cycle where experience affects future experience. Hence, a focus on experience is critical as it drives student engagement in learning and assessment in schools. Importantly, both qualities – the practical imperative and the focus on experience – not only apply to students' learning and assessment in classrooms but also to how teachers learn to assess; therefore, they operate as both the antecedent for and the consequence of teachers' assessment work.

Transdisciplinary

In contrast to viewing assessment literacy as *multidisciplinary*, in which each discipline takes up different perspectives, theoretical tenets, and research approaches to assessment literacy, or as strictly *interdisciplinary*, in which researchers from different disciplines cooperate to develop the theory, I assert that assessment literacy theory is best characterized as *transdisciplinary*. As Nicolescu (1996) characterized: 'transdisciplinarity concerns that which is at once between the disciplines, across the different disciplines, and beyond all disciplines. Its goal is the understanding of the present world, of which one of the imperatives is the unity of knowledge.' As such, a transdisciplinary understanding of assessment literacy avoids the reduction of assessment work to component parts and honors the complexity of factors by acknowledging what shapes teachers' approaches to assessment. Furthermore, such a view of assessment literacy theory aligns with its core tenets of pragmatism and complexity; as Lawrence and Després (2004:399) noted, transdisciplinary contributions 'tackle complexity in science' and are 'often action-oriented [and] frequently deal with real-world topics and generate knowledge that not only address societal problems but also contribute to their solution'.

Complexity

By now it is clear from the literature that assessment literacy is not a simple enterprise but rather a multi-dimensional professional capacity. This final tenet draws broadly on principles of complexity thinking (Cilliers 2010, Davis and Sumara 2006) to recognize that assessment literacy theory is greater than the sum of its parts and invariably complex. Assessment literacy is not simply knowledge nor beliefs nor feelings nor confidence (to draw on Looney et al's 2017 framework) but rather the combination and development of all these dimensions (and potentially others) to shape how a teacher approaches

assessment in their classroom. Unlike complicated theories that describe linear and predictable pathways and progressions, assessment literacy theory, at least based on what we know to date, is far more complex (similar to many other teaching capacities), with assessment literacy dimensions influencing teachers in different ways based on their background, experience, students, and context. At the heart of assessment literacy theory is teachers' capacities to negotiate the various knowledges (i.e., assessment, pedagogical, curricular, disciplinary, cultural, ethical, epistemological, etc.), relationships (i.e., students, peers, parents, administrators, etc.), and emotional spaces to enact and empower positive assessment practices and cultures in their classrooms. From such a theoretical basis stems the possibility for vast emergence and difference for how a teacher might take up assessment in their practice. This space of emergence enables assessment to move from the letter to the spirit in classrooms, giving teachers agency to integrate assessment into their pedagogies and in relation to the diversity of students and curricula they encounter. Such a complex orientation means that assessment literacy theory gives rise to the power of assessment in classrooms and not solely to specific assessment behaviours.

Concluding thoughts

Given the argument advanced in this chapter, where can we locate research that aims to root assessment and assessment theory within disciplinary contexts, for example, the numerous studies endeavoring to localize assessment literacy for language-learning contexts? I agree that discipline does shape how a teacher practices and negotiates assessment in their classroom – how could it not? However, drawing on assessment literacy theory, we know that discipline is not the primary factor that shapes assessment practice, but instead one contextual factor. If, as researchers, we begin to argue the need for assessment theories to be distinguished by disciplines – almost to the point where disciplinary assessment theories become their own theoretical domain – then the pursuit in front of us would be an endless one; not only could the case be made to distinguish theory by discipline but also by all matters of context – grade level, socioeconomic, cultures, etc. My argument here is not that assessment literacy should not be investigated within these nuanced contexts; to the contrary. My argument is that such investigations should not splinter the field into fragments, but rather, should ensure recursion back to general assessment literacy theory. A persistent question for all assessment literacy studies (within and across contexts) should be: *how does this research contribute to our evolving thinking about how teachers come to assess writ large?* The call here is to recognize assessment literacy as one transdisciplinary, pragmatic, and complex field, anchored by the unifying aim to understand the complex reality of how

teachers across contexts approach assessment in their varied classrooms to ultimately support student learning.

References

- Bennett, R E (2011) Formative assessment: A critical review, *Assessment in Education: Principles, Policy & Practice* 18 (1), 5–25.
- Cilliers, P (2010) The value of complexity, *Complicity: An International Journal of Complexity and Education* 7 (1), 39–42.
- Davis, B and Sumara, D J (2006) *Complexity and Education: Inquiries into Learning, Teaching, and Research*, New York: Routledge.
- Dewey, J (1938/1959) *Experience and Education*, New York: Macmillan.
- Hodgen, J and Marshall, B (2005) Assessment for learning in English and mathematics: A comparison, *The Curriculum Journal* 16 (2), 153–176.
- James, W (1907) *Pragmatism. A New Name for Some Old Ways of Thinking*, New York: Hackett Classics.
- Lan, C, and Fan, S (2019) Developing classroom-based language assessment literacy for in-service EFL teachers: The gaps, *Studies in Educational Evaluation* 61, 112–122.
- Lawrence, R J and Després, C (2004) Futures of transdisciplinarity, *Futures* 36, 397–405.
- Looney, A, Cumming, J, van Der Kleij, F and Harris, K (2017) Reconceptualising the role of teachers as assessors: Teacher assessment identity, *Assessment in Education: Principles, Policy & Practice* 25 (5), 1–26.
- Marshall, B and Drummond, M J (2006) How teachers engage with assessment for learning: Lessons from the classroom, *Research Papers in Education* 21 (2), 133–149.
- Nicolescu, B (1996) *La Transdisciplinarité, Manifeste*, Paris: Le Rocher.
- Ormerod, R (2006) The history and ideas of pragmatism, *Journal of the Operational Research Society* 57, 892–909.
- Pastore, S and Andrade, H L (2019) Teacher assessment literacy: A three-dimensional model, *Teaching and Teacher Education* 84, 128–138.
- Ramadier, T (2004) Transdisciplinarity and its challenges: the case of urban studies, *Futures* 36, 423–439.
- Willis, J, Adie, L and Klenowski, V (2013) Conceptualising teachers' assessment literacies in an era of curriculum and assessment reform, *The Australian Educational Researcher* 40 (2), 241–256.
- Xu, Y and Brown, G T L (2016) Teacher assessment literacy in practice: A reconceptualization, *Teaching and Teacher Education* 58, 149–162.

