Language Test Validation in a Digital Age

Also in this series:

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

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

On Topic Validity in Speaking Tests
Nahal Khabbazbashi

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 and Lynda Taylor

Language Assessment Literacy and Competence Volume 2: Case Studies from Around the World
Edited by Beverly Baker and Lynda Taylor

Language Test Validation in a Digital Age

Edited by
Guoxing Yu
University of Bristol, UK
and
Jing Xu
Cambridge University Press & Assessment, UK

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

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

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

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

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

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

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

www.cambridge.org
Information on this title: www.cambridge.org/9781108931908

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 978-1-108-93190-8

Contents

Part 2
Using technology to enhance language assessment

Series Editors' note

This SiLT volume has its origins in 2016 when Professor Stephen Bax was invited by the Series Editors (Saville and Weir) to produce a SiLT volume on the use of technology in language assessment. By then Stephen already enjoyed a well-deserved reputation in this field and he was pleased to accept the invitation. He was making good progress in 2017 and, as Jing Xu and Guoxing Yu explain in their Preface to this volume, he had invited Jing to collaborate with him as co-editor. Unfortunately, further progress and collaboration were abruptly interrupted in December that year by Stephen's untimely and unexpected death.

After a period of reflection in 2018, the Series Editors decided that work on the volume should continue as a mark of respect to Stephen. We are pleased to say that Guoxing accepted the invitation to join Jing as co-editor and together they undertook the task of completing the planned publication. We are grateful to them for their expertise, commitment and perseverance. As they explain in the Preface, there were several more bumps along the way in finishing off the work that Stephen had started.

Stephen's collaboration with the Series Editors had begun while he was working with Professor Cyril Weir and his team at the Centre for Research in English Language Learning and Assessment (CRELLA), the University of Bedfordshire between 2009 and 2016. He participated in several research projects involving both CRELLA and Cambridge (then called Cambridge English Language Assessment) in using eye-tracking techniques. He also started collaborating with the Association of Language Testers in Europe (ALTE) in 2016 to deliver seminars and training courses, including a bespoke summer course on uses of technology in language test production and validation. These experiences certainly influenced the approach he took to planning the volume.

However, Stephen's broad interest in language and technology was evident much earlier in his career. In the early 2000s, he developed an interest in Computer Assisted Language Learning (CALL) while at Canterbury Christ Church University, where he established CRADLE: the Canterbury Centre for Research, Assessment and Development in Language Education. In that period, he was awarded a prize by Elsevier for his article in *System* entitled 'CALL – Past, Present and Future'. A few years later he developed Text Inspector, an online tool to analyse the difficulty of written texts, which won the British Council award for Best Digital Innovation Programme.

While at CRELLA, he also collaborated with the Cambridge team working on the English Profile Programme and made Text Inspector available to produce a tool for analysing texts in terms of the English Vocabulary Profile.

At the time of Stephen's death, he had just moved to a professorial post in the Open University with the possibility to extend his academic interests in that new role.

The co-editors are to be commended for bringing together an excellent collection of chapters representing many contemporary voices from a range of sectoral and geographical contexts, in keeping with Stephen's original vision. Our thanks go to the authors of those chapters and to the many reviewers who helped in the editorial process. The finished volume represents a fitting tribute to Stephen's work and his legacy. As such it is also a timely addition to the SiLT series and to the field in general.

Nick Saville
Lynda Taylor

May 2024

Preface

This Studies in Language Testing (SiLT) volume was originally proposed by Professor Stephen Bax, an internationally recognised scholar for his work on eye tracking, learners' cognitive processes in reading and reading assessment (Thomas and Motteram 2019). In September 2016, Stephen was invited by Cambridge (then called Cambridge English Language Assessment) to teach a one-week summer course for the Association of Language Testers in Europe (ALTE), on the topic of 'Technology in language test production and validation'. The course covered six sub-topics, including: the role of technology in the ALTE framework; online testing, specifications, and principles; constructing tests with technology; technology in building a validation argument; technology in scoring and marking; and technology in classroom assessment. Jing, who joined Cambridge in November 2015, was invited by Stephen to deliver a guest lecture on automated assessment of speaking and writing. Stephen showed great interest in the use of machine learning technology to mark constructed responses.

In March 2017 Stephen invited Jing to a research seminar on 'The use of technology in reading/writing assessment research', jointly hosted by the Centre for Research in English Language Learning and Assessment (CRELLA) and the School of Computer Science at the University of Bedfordshire. During a 'CRELLA walk' after lunch on the Putteridge Bury campus, Stephen discussed with Jing the idea of co-editing a SiLT volume focusing on technology, as he saw a rapidly increasing impact of technology use on language assessment. The book proposal was welcomed by Professor Cyril Weir and Dr Nick Saville, then the Series Editors of SiLT. It was deemed as a timely contribution to the digital transformation of English language assessment taking place in Cambridge (Chan, Bax and Weir 2018, Saville 2017).

This book project, however, experienced unexpected turbulence from the beginning. Stephen's health degenerated and he sadly passed away in November 2017. Guoxing stepped into Stephen's role in January 2018. Cyril's passing in September 2018 came as a second blow. Professor Lynda Taylor kindly offered to help. Then came the COVID-19 pandemic in which a lot of work slowed down, and a few authors had to withdraw due to their personal circumstances.

We are profoundly grateful to all those who contributed to this volume for their wholehearted support, persistence, and patience. The authors

of this volume were marvellous in that many of them not only wrote their own chapters but also helped review chapters written by others. We are also indebted to the following non-author reviewers who offered valuable feedback and critique on the early drafts of each chapter.

- Aaron Olaf Batty, Keio University
- Andrea Révész, University College London
- Benjamin Kremmel, University of Innsbruck
- Carol Chapelle, Iowa State University
- Dan Douglas, Iowa State University
- Erik Voss, Teachers College, Columbia University
- Gad Lim, Cambridge Boxhill Language Assessment
- Hee Sung (Grace) Jun, Seoul National University
- Hye-won Lee, Cambridge University Press & Assessment
- James Wollack, University of Wisconsin-Madison
- Jason Fan, The University of Melbourne
- Jinrong Li, Georgia Southern University
- John Pill, Lancaster University
- Kathrin Eberharter, University of Innsbruck
- Nicola Latimer, University of Bedfordshire
- Philippa Howard, University of Bristol
- Shinhye Lee, Educational Testing Service
- Sonja Zimmermann, TestDaF-Institut
- Tineke Brunfaut, Lancaster University
- Vahid Aryadoust, Nanyang Technological University Singapore
- Xun He, Bournemouth University
- Xun Yan, University of Illinois at Urbana-Champaign
- Zhi Li, University of Saskatchewan

We would like to express our profound gratitude to Dr Nick Saville and Professor Lynda Taylor who steered the direction and scope of this project. We are deeply grateful to John Savage, the publications assistant of this project, for his professional work in proofreading, copy-editing, and communicating with the authors and the publisher.

Finally, we would like to pay tribute to Professor Stephen Bax and Professor Cyril Weir for their distinguished achievements in advancing the field of language assessment.

Jing Xu
Guoxing Yu

References

Chan, S, Bax, S and Weir, C J (2018) Researching the comparability of paper-based and computer-based delivery in a high-stakes writing test, *Assessing Writing* 36, 32–48.

Saville, N (2017) Digital assessment, in Carrier, M, Damerow, R M and Bailey, K M (Eds) *Digital Language Learning and Teaching: Research, Theory, and Practice*, New York: Routledge, 198–207.

Thomas, M and Motteram, G (2019) Editorial for the special edition commemorating the work of Stephen Bax, *System* 83, 1–3.

Notes on contributors

Vivien Berry was Senior Researcher of English Language Assessment at the British Council when the studies in this volume were conducted. Her research interests include the impact of individual learner characteristics on performance in oral tests, the use of technology in oral language assessment, and teachers' assessment literacy. Her publications include authored and edited books, chapters in several titles in the Cambridge University Press & Assessment SiLT series, and articles in peer-reviewed international journals.

Sathena Chan is an Associate Professor in Language Assessment at the Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire. She conducts research on different aspects of large-scale language assessment from construct definition, test design, rating scales to validation. Her work explores the dynamic relations between multimodal input-based tasks, test-takers' cognitive processes and levels of performance. She has published widely in peer-reviewed journals such as *Language Testing*, *Assessing Writing* and *System*. Her books include *Defining Integrated Reading-into-Writing Constructs: Evidence at the B2-C1 Interface* (2018) and *Research and Practice in Assessing Academic Reading: The Case of IELTS* (2019), both co-published by Cambridge Assessment English and Cambridge University Press (now Cambridge University Press & Assessment).

Elena Cotos holds a dual appointment at Iowa State University – as Associate Professor of Applied Linguistics in the Department of English and as Associate Dean in the Graduate College. In her faculty role, she leads a research agenda that bridges corpus-based genre analysis, genre-based automated writing evaluation, and writing pedagogy, which has informed the development of educational programs, curricula, instructional materials, and computer-assisted learning and assessment tools focused on English for academic purposes. Her work on the Research Writing Tutor is described in her first book, *Genre-Based Automated Writing Evaluation for L2 Research Writing* (Palgrave Macmillan, 2014) and in her contributions to numerous edited volumes. Her research has also been published in *Journal of English for Academic Purposes*, *English for Specific Purposes*, *Language Learning and Technology*, *Language Testing*, *ReCALL*, *CALICO Journal*, *Journal of Writing Research*, *Writing and Pedagogy*, *Journal of Learning Analytics*,

amongst others. She currently serves as an Associate Editor for the journal *English for Specific Purposes*. In the Graduate College, she is the founding Director of the Center for Communication Excellence and oversees the professional development programs. She is the institutional lead for global and massive open online courses on developing and teaching academic writing, and on establishing academic writing centres at international higher education institutions, offered in partnership with the US Department of State's Online Professional English network and FHI 360.

Maggie Dunlop is a Senior Statistical Research Analyst at the Education Quality and Accountability Office, an agency of the Government of Ontario, Canada. She completed her doctoral studies at the Ontario Institute for Studies in Education, University of Toronto, specialising in language assessment and educational measurement. Her dissertation focused on how adult immigrant English language learners in Canada process and use feedback on their English reading skills. She previously completed her Master's studies at the George Washington University, Washington, D.C., specialising in language education policy and programme evaluation for education initiatives. She has over 16 years' experience working internationally, including in East Asia, Europe, North America and other regions, on programme evaluation and research initiatives that have mainly supported improving language and literacy learning outcomes for multilingual learners and learners in resource-restricted contexts. Her main areas of technical expertise are programme evaluation, quantitative/mixed methods of inquiry, assessment development, and psychometrics.

Evelina Galaczi is Director of Research-English at Cambridge University Press & Assessment. She has worked in English language education for over 30 years, and her current work focuses on the challenges – and exciting opportunities – of using AI in language learning, teaching and assessment. She serves as a Trustee for the International Research Foundation for English Language Education and is co-editor of the journal *Language Assessment Quarterly*.

Mark Gales is a Professor of Information Engineering and a College Lecturer and Official Fellow of Emmanuel College, University of Cambridge. His research interests include automatic spoken language assessment; uncertainty, generalisation and interpretability in deep learning systems; and large vocabulary automatic speech recognition. He led the development of the automated spoken language assessment technology under the Automatic Language Teaching and Assessment (ALTA) Institute, University of Cambridge, that is deployed in the Linguaskill Speaking

test. He was the Principal Investigator for the ALTA Spoken Language Processing (SLP) Technology Project from 2013–2021 and continues as a Technical Advisor. He is a Fellow of the Institute of Electrical and Electronics Engineers (IEEE) and International Speech Communication Association (ISCA), and is on the Editorial Board of *Computer Speech and Language*. He has over 400 publications in international peer-reviewed journals and conferences and an h-index of 66. He has been awarded a number of paper awards, including a 1997 IEEE Young Author Paper Award and a 2002 IEEE Best Paper Award.

Chihiro Inoue is Associate Professor in Language Assessment at the Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire. She specialises in the assessment of L2 speaking, with particular interest in task design features and learner language, test accommodations, and assessment literacy, and she has carried out numerous funded local and international test development and validation projects around the world. Her publications have appeared in peer-reviewed journals such as *Language Assessment Quarterly, Assessing Writing, Assessment in Education* and *Language Learning Journal.*

Edmund Jones is a Senior Research Manager at Cambridge University Press & Assessment, working on research related to computer-based tests, automated assessment, psychometrics, and detection of malpractice. He holds a PhD in computational statistics and previously worked at the University of Cambridge on statistical modelling of cardiovascular disease. His work has been published in *Medical Decision Making*, *The Lancet* and *Assessment in Education: Principles, Policy & Practice*.

Kate Knill is a Principal Research Associate at the Department of Engineering and the Automatic Language Teaching and Assessment (ALTA) Institute, University of Cambridge. Since 2021 she has been the Principal Investigator for the ALTA Spoken Language Processing (SLP) Technology Project. She is a co-contributor to the automated spoken language assessment technology transferred from this project as described in the UCLES 2020 Linguaskill Research Report *Linguaskill – building a validity argument for the Speaking test*, of which she is a co-author. She has over 30 years' experience in spoken language processing in industry and academia. She has led and contributed to the development of automatic speech recognition and text-to-speech synthesis systems for over 50 languages and dialects. Her current research focus is on education – computer-aided assessment and learning of English speaking skills for L2 learners – and medical fields – screening for speech and language disorders. She is a Fellow of the International Speech Communication Association (ISCA).

Suh Keong Kwon is a Professor in the Department of English Education at Chinju National University of Education in the Republic of Korea. He holds a PhD in Education (Language Testing) from the University of Bristol. His research interests include language assessment, technology-assisted language learning and testing, and eye-tracking methodology. His work has been published in peer-reviewed journals such as *Language Testing, Language Testing in Asia*, *Language Learning & Technology, System* and *Assessment in Education: Principles, Policy & Practice.*

Shinhye Lee obtained her PhD degree in Second Language Studies at Michigan State University, with a specialisation in Language Assessment. Over the recent years, she has worked as a language test developer and assessment researcher at Michigan State University, at the University of Chicago, and most recently, at Educational Testing Service. Her academic publications have appeared in journals such as *Language Testing, TESOL Quarterly* and *Computer Assisted Language Learning*.

Stephanie Link, PhD, is an Associate Professor of Applied Linguistics and Technology at Oklahoma State University, where she is the Director of Graduate Studies. She also serves as the Series Editor for the Advances in CALL Research and Practice Book series and Book Review Editor for the journal *English for Specific Purposes*. Her research uses social and cognitive approaches to second language teaching/learning and learner-centred design, and focuses on the development and appraisal of automated evaluation tools and intelligent tutoring systems for academic and scientific writing. Her work is in peer-reviewed journals, including *CALICO Journal, Computer Assisted Language Learning, Journal of Second Language Writing* and *Language Learning and Technology*. Her co-edited book *Assessment Across Online Language Education* can be found through Equinox Publishing.

Fumiyo Nakatsuhara is Professor of Language Assessment and Deputy Director of the Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire. Her research interests include the development and validation of speaking tests, assessment of interactional competence, and educational reform through assessment. She has led numerous international projects, working with ministries, universities and examination boards

Mikako Nishikawa is an Associate Professor in the School of Global Humanities and Social Sciences at Nagasaki University. She completed her PhD in Applied Linguistics at the University of Bristol. Her research interests include language assessment and teaching, second language acquisition (SLA) and curriculum design. She is currently leading two research projects

funded by the Japan Society for the Promotion of Science (JSPS). Her recent projects include the 'Eye-tracking Study: Investigating the underlying constructs of the listening-to-summarize English tasks (JSPS #1K00733)' and 'An Eye-Tracking Study: Exploring Integrated Reading Tasks in the New Format of the English Common Test for Japanese University Admissions (JSPS# 24K04032).'

Nathaniel Owen is a Senior Research and Validation Manager at Oxford University Press. He was formerly a postdoctoral Research Associate at the Open University, where the research cited in his chapter was completed. He holds a PhD in language testing from the University of Leicester specialising in L2 reading processes. In addition to reading processes, his research interests include the interface of language testing and technology, big data analytics, the use of language tests in English-medium instruction contexts, research methods and widening participation in higher education. He has conducted funded research projects for examination boards in the UK and the US. His most recent publications include the two monographs *Researching Academic Reading in Two Contrasting English as a Medium of Instruction Contexts at a University Level* (2021) and *Researching lexical thresholds and lexical profiles across the CEFR assessed in the Aptis test* (2021), published by Educational Testing Service and the British Council, respectively.

Carla Pastorino-Campos is a Principal Research Manager at Cambridge University Press & Assessment. Her research interests include the cognitive, affective and psychological aspects of L2 learning and assessment, the effect of individual differences in L2 acquisition and performance, and the application of technology to language learning and research. She is also interested in the ethical aspects of language assessment. She has conducted research projects on topics ranging from the effectiveness of access arrangements for test-takers with literacy difficulties to the predictive validity of high-stakes tests.

Elaine Schmidt is a Senior Research Manager at Cambridge University Press & Assessment. Elaine's research focuses on cognitive aspects of language processing and learning using eye tracking and electroencephalography (EEG). She obtained her PhD in phonetics and bilingual language acquisition from the University of Cambridge. After her PhD she worked on cognitive processes of L1 and L2 speech perception in Sydney, Australia, before she moved back to the Linguistics Department at the University of Cambridge. A few years later she then decided to combine her research with more practical applications and joined Cambridge University Press & Assessment, where she brings her expertise in speech production and perception, eye tracking and EEG in second language learning to an assessment context.

Ruslan Suvorov is an Associate Professor in Applied Linguistics at the University of Western Ontario, Canada, where he teaches courses in second language assessment and computer-assisted language learning (CALL). His research interests lie at the intersection of language testing and assessment, CALL, and instructional technology and design, with a particular focus on second language listening. He has given numerous presentations and workshops at various regional, national, and international conferences and published in *Language Testing, Language Learning & Technology, International Journal of Listening, CALICO Journal,* as well as in edited volumes, conference proceedings, encyclopaedias and research reports. He is a co-author of *Blended language program evaluation* (Palgrave Macmillan, 2016).

Paula Winke is a Professor at the Department of Linguistics and Languages at Michigan State University, where she also directs the Second Language Studies PhD program. She is the co-editor (with Professor Luke Harding) of the journal *Language Testing*. She researches language assessment methods, language test score uses, and how to best measure latent traits for second language acquisition (SLA) research. She is on the editorial board of *The Modern Language Journal*, and is currently serving on the 'Task Force on the Future of Language Testing' with the US Foreign Service Institute. Her work received the 2008 Article of the Year Award from *CALICO Journal*, the 2014 Distinguished Research Award from TESOL International, and the 2020 Research Article of the Year Award from the American Association of Applied Linguistics. Recent publications include *The Routledge Handbook of SLA and Language Testing* (2021), co-edited with Professor Tineke Brunfaut, and *A Principled Approach to Language Assessment: Considerations for the US Foreign Service Institute* (2020), co-written as a National Academy of Sciences ad hoc committee member and published by The National Academies Press.

Jing Xu is Head of Propositions Research at Cambridge University Press & Assessment. His research interests are in the application of technologies, particularly artificial intelligence, in language assessment and learning and the related validity issues. His current work focuses on automated marking of L2 speaking and writing performance in high-stakes language assessment. He was the winner (2017) of the Jacqueline A. Ross Dissertation Award and the winner (2010) and runner-up (2021) of the ILTA Best Article Award. He received his PhD in Applied Linguistics and Technology from Iowa State University. Website: www.languagesciences.cam.ac.uk/staff/dr-jing-xu-0

Xun Yan is an associate professor of Linguistics and Educational Psychology at the University of Illinois Urbana-Champaign (UIUC), where he also

directs the undergraduate program in Linguistics + TESOL and the university-level English Placement Test. His research interests include speaking and writing assessment, psycholinguistic approaches to language testing, and language assessment literacy. His work has been published in peer-reviewed journals such as *Language Testing*, *Language Assessment Quarterly*, *Assessing Writing*, *Applied Linguistics*, *TESOL Quarterly*, *Journal of Second Language Writing*, *System* and *Foreign Language Annals*.

Guoxing Yu is Professor of Language Assessment at the University of Bristol. His research interests include language assessment, applied linguistics (language teaching and acquisition), assessment of learning power, and educational assessment policies. He has published in *Applied Linguistics, Applied Linguistics Review, Assessing Writing, Assessment in Education, Educational Research, Interactive Learning Environments, International Journal of Bilingual Education and Bilingualism, International Journal of Listening, Journal of Computers in Education, Journal of Eye Movement Research, Language Assessment Quarterly, Language Learning & Technology, Language Teaching Research, Language Testing, Reading and Writing, System*, among others. Website: research-information.bris.ac.uk/en/persons/guoxing-yu

1 Introduction

Guoxing Yu
University of Bristol, UK

Jing Xu
Cambridge University Press & Assessment, UK

The beginning of the 2020s saw massive disruptions caused by the coronavirus pandemic. Social restrictions aiming to prevent the spread of the virus made remote working and learning more prevalent and acceptable. The ways people communicate have adapted accordingly, with virtual or hybrid meetings becoming a norm. The pandemic also exerted a profound impact on the language testing industry and altered its landscape. Test centres had to be closed; millions of in-person tests were cancelled or rescheduled (Clark, Spiby and Tasviri 2021). Both global and local language test providers had to put forward alternative testing methods facilitated by technology to allow candidates to sit tests in a safe environment (Ockey 2021). In this sense, the pandemic accelerated the digitalisation of language assessment.

Technology has been deemed as a resource for supporting, improving, and even revolutionising language assessment (Chapelle and Voss 2016, Sawaki 2012, Schmidgall and Powers 2017). Various terms were used to describe tests mediated by computer technology, such as computer-based, computer-aided, computer-assisted, computer-enhanced, computer-mediated, computer-supported, and computer-integrated (Yu and Zhang 2017). The late 1990s and early 2000s saw a move to computer-delivered language tests which aimed to improve test efficiency and accessibility. A computer-based version of the Test of English as a Foreign Language (TOEFL CBT) was introduced in 1998 and then followed by the launch of the internet-based version of the test (TOEFL iBT) in 2005 (Alderson 2009, Chapelle, Enright and Jamieson (Eds) 2008). Meanwhile the computer-adaptive testing (CAT) technique featuring item banking, automatic item selection and automated scoring of objective items was applied to language assessments such as the Business Language Testing Service (BULATS) and CommuniCAT (Chalhoub-Deville and Deville 1999, Geranpayeh 2001). Computer technology offered convenience to test administration by allowing on-demand testing, onscreen marking, and online training and management of examiners (ffrench, Bridges and Beresford-Knox 2012).

The second decade of the 21st century saw a stronger presence of artificial intelligence (AI) in language assessment, with a primary goal to improve

efficiency and save human resource in test construction, delivery, and administration. A special issue of *Language Testing* (Xi 2010) focused on automated scoring and feedback systems. This resonated with the increased use of machine scoring in high-stakes English language tests including TOEFL iBT and Versant (Bernstein, Van Moere and Cheng 2010, Enright and Quinlan 2010) as well as low-stakes formative assessments such as TOEFL iBT Speaking Practice (Xi 2008, Xi, Schmidgall and Wang 2016), Criterion (Burstein, Chodorow and Leacock 2004) and Cambridge English Write & Improve (Briscoe, Medlock and Andersen 2010, Yannakoudakis, Øistein, Geranpayeh, Briscoe and Nicholls 2018). The Duolingo English Test (DET), launched in 2016, features automated scoring, automated generation of test questions, and automated detection of suspicious cheating behaviours using AI algorithms (LaFlair et al 2022). Recently, research into intelligent dialogue systems for eliciting learner speech in paired oral tasks has also shown great promise in mitigating the challenges of delivering conversational speaking assessment on a large scale (Karatay 2022, Ockey and Chukharev-Hudilainen 2021).

While efficiency is important, it may not be a 'sufficient reason' for integrating technology into language assessment (Chapelle and Douglas 2006:44). Researchers have argued for this integration to make assessment tasks more authentic and relevant, on the basis that human communication has become more digital and technology-reliant. Douglas (2013:2) urged test developers to define the language construct 'to include appropriate technology in light of the target situation and test purpose'. Thus, if language use in the target situations requires computer technology, then the target construct is expected to reflect this (see also Yu and Zhang 2017 on their recommended term 'computer-integrated language testing'). Alderson (2009) critiqued TOEFL iBT for not assessing a candidate's ability to revise, refine and polish writing. Such revision processes are commonly supported by current technology use, for example, AI-powered assistive tools such as Grammarly. In their scoping review, Khabbazbashi, Chan and Clark (2023) observed the increasing use of digital technologies in higher education and suggested including technology-based multimodal communication in assessing English for Academic Purposes (EAP). Likewise, Chapelle and Voss (2016) discussed the potential of using tasks in an asynchronous online discussion forum as a form of assessment in an EAP course. They believed that learner participation in such online forums had become a common activity in language education. Suvorov and He (2022), Suvorov (**Chapter 3** of this volume), and Kwon (**Chapter 4** of this volume) supported multimodal second language (L2) L2 listening assessments on the basis that modern-day listening often requires processing both auditory and visual information. Nakatsuhara, Inoue, Berry and Galaczi (2017) and Inoue, Nakatsuhara, Berry and Galaczi (**Chapter 10** of this volume) explored speaking assessment

using video-conferencing technology given that virtual or hybrid meetings had become widespread. Thus, if language use in the target situations is mediated by technology, it seems plausible to reflect this interdependence between language and technology in the test construct.

Besides modernisation of the test construct, researchers have been seeking 'true innovations' that would fundamentally change the ways in which language assessment is done (Sawaki 2012:429). It is believed that technology holds the promise of providing better and more sophisticated means for measuring language abilities (Chapelle and Douglas 2006) and plays a pivotal role in integrating language assessment and learning (Chapelle and Voss 2016, Jones and Saville 2016). For example, technology has offered assessment researchers new opportunities to advance their knowledge about language constructs. Not satisfied with the psychometric evidence of test scores, researchers have become keener to explore test-takers' cognitive processes underlying their responses. For example, Winke, Yan and Lee (**Chapter 2** of this volume) investigated test-takers' eye-movement data to better understand the construct of an L2 cloze test. Schmidt and Pastorino-Campos (**Chapter 8** of this volume) discussed the potential of electroencephalography (EEG) for studying the construct of reading. Suvorov (**Chapter 3** of this volume) used eye-tracking methods to explore the processes and strategies involved in test-takers' responses to multimedia-enhanced listening items. Research in this area is crucial for refining cognitive theories of language processing and defining language constructs more precisely.

While modern AI technologies seem to have the potential to increase the accessibility and efficiency of language assessment, they also present the language testing community with unprecedented challenges. There have been serious concerns over construct underrepresentation or misrepresentation caused by the use of automated scoring technology in assessment and its potential negative impact on language teaching and learning (e.g., Chun 2006, 2008, Xi 2010, 2022, Xu 2015). A burning issue that researchers currently encounter is the opacity of scoring algorithms, particularly those developed based on neural network models (Khabbazbashi, Xu and Galaczi 2021). The unexplainable nature of these models leads to increased difficulty in test score interpretation. Additionally, the score or feedback given by an AI system is not always reliable or accurate (Link, **Chapter 11** of this volume; Liu and Yu 2022). According to Leslie (2019), the misuse or poor design of AI systems may cause a range of individual and societal harms as these systems may 'reproduce, reinforce, and amplify' the patterns of bias and discrimination that exist in the data used to train them (2019:4). Digital-first assessments may also face serious challenges in capturing gaming behaviours, such as fraudulent responses generated by Large Language Models (LLMs) (LaFlair et al 2022, Xi et al 2016). Furthermore, technology-enhanced language assessment may exacerbate the digital divide. Many testing organisations

prepare both paper-and-pencil and digital modes of assessments. However, it is generally difficult to ensure the comparability of the language constructs across the two modes – technology-mediated language use may require additional 'technological and strategic competence' on top of linguistic competence (Chapelle and Douglas 2006:107).

To understand the role of technology in language assessment, we adopted Weir's (2005) socio-cognitive framework for language test validation. Technology intersects with all five aspects of validity in this framework. *Context validity* is concerned with the degree of representativeness of the test tasks to the target language use situations. Thus, if the targeted language abilities are about technology-mediated communication, the test tasks are expected to reflect these characteristics. *Cognitive validity* is about the underlying processes of test behaviours and the interpretability of such behaviours based on relevant construct theories. As discussed above, technologies such as eye tracking and EEG can provide new means to investigate the cognitive processes involved during test-taking. The new tools can help collect empirical evidence to support, refine or even refute existing theories. *Scoring validity* refers to the reliability of scoring process and test scores. The use of automated scoring techniques is closely related to this aspect of validity. *Criterion-related validity* is about the relationship between test performance and some external criteria of language performance. For example, if a test claims to assess multi-modal communication skills, we expect to see a high correlation between candidates' test scores and their language performance in a multimodal environment. *Consequential validity* is about the social consequences of assessment. If the technology used in assessment led to inequality or discrimination, or teaching and learning of knowledge and skills that deviate from the language abilities that address learners' real-world communication needs, then the consequential validity of the assessment would be undermined.

Summary of the chapters

This edited volume consists of two parts. The eight chapters in Part 1 report studies that have used technology (e.g., eye tracking) to *validate* language assessment tasks (and feedback from assessment), and the five chapters in Part 2 report studies that have used technology to *enhance* language assessment task development, delivery and test-taker experiences. Below we present a summary of these chapters.

Part 1: Using technology to validate language assessment

Winke, Yan and Lee (**Chapter 2**) used eye tracking (Tobii TX300) as the main data collection tool to explore the construct of an online French cloze test, a

conceptual replication study of Tremblay's (2011) original research on this type of test. Their first research question examined whether test-takers' eye movements on the cloze test demonstrated a largely linear and sequential movement pattern of reading comprehension; and the second question aimed to understand test-takers' perceptions on the difficulty and appropriateness of the cloze test as a measure of their general language proficiency. Using mixed research methods (eye movements and interviews with 22 test-takers) and from the perspective of test score interpretations, their study aimed to provide new insights into whether the original claims made by Tremblay that the French cloze test can assess the general language proficiency of any French learner and can identify learners according to the American Council on the Teaching of Foreign Languages (ACTFL) proficiency scale are valid. The findings of the study broadly supported Tremblay's argument that the French cloze test can discriminate low-level from high-level learners well, but the study also found that it was more difficult to accurately split adjacent-performing individuals. The cloze test was also found to be too cognitively demanding and therefore not appropriate for lower-level learners. Therefore, Winke et al concluded that the cloze test cannot adequately assess learners' French ability. This chapter is an excellent example of a mixed research methods study employing eye tracking as the main tool for collecting an online, introspective and cognitive test-taking process with the supplement of retrospective test-takers' interviews, so that the construct of a language assessment task could be investigated from multiple perspectives and thus better understood.

Suvorov (**Chapter 3**) reported two eye-tracking studies on second language learners' cognitive processes and test-taking strategies in listening comprehension tests. Study 1 explored the viewing behaviours of test-takers during their interactions with two types of visuals (content videos and context videos) in an online video-based listening test. Two research questions were asked: (1) To what extent do L2 test-takers watch context videos differently from content videos? (2) What aspects of visual information in the two video types do L2 test-takers find helpful and/or distracting for their listening comprehension and test performance? EyeTech Vision Tracker 2 (80 hertz (Hz) data sampling rate) was used to record the eye movements of 33 test-takers while they were completing the academic listening test, which consisted of short video clips of authentic academic lectures as input. Test-takers were allowed to take notes on paper during the test. After completing the test and cued by their recorded eye movements, each test-taker retrospectively reported on their viewing behaviours and how they used visual information from the videos. Study 2 examined test-taking strategies used by test-takers when completing three types of listening items (discrete dialogue, dialogic listening, and monologic listening) delivered by the computer as audio files. It also reported the effect of test-wiseness strategies on their test scores.

A Gazepoint GP3 (60 Hz) eye tracker was used to record the eye movements of 15 test-takers while they were completing the listening test. As in Study 1, test-takers were allowed to take notes on paper during the test. After completing each set of items (again cued by their recorded eye movements), each test-taker retrospectively reported on their use of strategies for answering each item and the reason why a particular option was chosen. The two studies demonstrated how eye-tracking technologies could be used to collect data on test-takers' response processes as cognitive validity evidence of the tasks. The two studies also demonstrated how different methods could be used to analyse eye-movement data and highlighted, as Winke et al (Chapter 2) did, the importance of incorporating and triangulating data from other sources.

Kwon (**Chapter 4**) compared the test-taking processes and viewing behaviours of Korean learners of English in secondary schools while they were completing two different modes of a listening comprehension test (audio-only and video-based) with the same audio content. The study examined how the inclusion of visual cues in listening comprehension tests would affect test-taking processes and performance. The two groups of test-takers (117 students in total, randomly assigned to one of the two test conditions) were compared on their test scores and eye movements, which were recorded in Tobii Studio using a Tobii X2-60 eye tracker. Kwon reported the differences between the audio-only and video-based groups in three fixation measurements (fixation count, fixation duration, and proportion of total fixation duration) on several areas of interest (e.g., speakers, PowerPoint slides, stem, key option, distractor options) of the listening comprehension test items. It was found that the video group achieved statistically significantly higher, though not substantial, test scores than the audio group. The eye-movement data showed that students in the audio group looked at the questions and options for significantly longer and more frequently than students in the video group. He attributed the significant difference between the two groups to the additional visual input that the video group was given, as the video group students spent more than half of the total test time looking at the speakers and PowerPoint slides. Though the inclusion of visual input in the listening test led to substantial changes in test-takers' viewing behaviours, as expected, it did not lead to substantial increase in test scores. From the perspectives of task authenticity, Kwon argued for presenting visual input in second language listening tests.

Owen (**Chapter 5**) reported what item completion processes test-takers went through in response to questions that require either basic comprehension or inferencing. Fourteen test-takers, who were at about the C1 level of the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001) completed a TOEFL iBT reading task. He used a Tobii Pro X3-120 eye tracker to record test-takers' eye movements

during the test. Each participant was then asked to elaborate on their item completion processes at the stimulated recall interviews. Based on the eye-movement metrics (e.g., the total number of saccades, backward sweeps, and number of fixations) and the stimulated recall interview data, Owen reported that test-takers used a form of careful local reading for both inferencing and basic comprehension items, rather than expeditious or global reading. The careful local reading involved a quick decision about which part of the text to read or re-read based on an expeditious word spot strategy. Although inferencing items engaged test-takers significantly more in backtracking (backward sweeps) than basic comprehension items, there were no significant differences in the number of fixations or saccades between the two item types. Owen argued that this finding indicated that the greater cognitive load and complexity of inferencing items stimulated more localised re-reading than basic comprehension items. Owen further elaborated on the utility and challenges of eye-tracking technology for test validation. He suggested that there were three challenges in using eye tracking as a data collection method in researching reading comprehension. Firstly, the domain relevance or comparability between the reading tasks in an eye-tracking experiment and those in the target language use domain. Secondly, the limitations of eye movement for examining or measuring higher-order processes which may not be observable. What is observable does not offer direct evidence for interpreting higher-order processes beyond lexical level. Thirdly, the difficulty in generalising the eye-movement data beyond individuals, as well as the difficulty in replicating eye-tracking research.

Nishikawa and Yu (**Chapter 6**) reported an eye-tracking study on test-takers' cognitive processes when completing graph-based integrated writing tasks. These tasks have multiple texts and graphs as input which are thematically linked to each other; test-takers were asked to summarise the key information from these sources. The study collected data from 42 Japanese secondary school students. It used a sequential explanatory design, including eye tracking, questionnaires and focus group discussions as its main data collection tools. The study aimed to explore the relationships between test-takers' performances on the writing tasks and the cognitive processes involved. The paper reported the correlations between the participants' fixation duration on various areas of the task prompts (e.g., the graphs, different parts of the texts, areas where the participants typed in their responses) and their writing performance in five sub-scores (i.e., main idea, coherence, cohesion, lexical range and accuracy, and grammatical range and accuracy) as well as the total score. It is worth noting that the fixation duration data in this paper was reported as a ratio rather than a raw or absolute value because the total fixation duration on a task varied a lot among the participants. A ratio of fixation duration on a particular area could better demonstrate a participant's attention to the area in comparison to another

participant who might have a substantially higher or lower total duration in the task. The data from the questionnaire completed by 37 participants, and the focus group discussions with 24 participants on the cognitive processes, provided further evidence on the potential effects of the participants' English language proficiency on how they engaged with the source texts and the graphs. Nishikawa and Yu discussed the implications of their findings from the perspectives of designing writing assessment tasks and conducting eye-tracking research. They pointed out the necessity of triangulating data from different sources and acknowledged a theoretical challenge that perhaps all eye-tracking studies in the field of language testing encounter. The challenge is to establish theoretically that longer or shorter fixation on a particular area of the task prompt would necessarily lead to worse or better test scores, and whether a longer fixation on a particular area of task prompts would mean that the participant is interested in or struggling with understanding that feature of a task prompt.

Dunlop (**Chapter 7**) compared three different methods of measuring the attention that adult immigrant English language learners paid to the feedback reports on their reading test performance. The three methods were eye tracking immediately followed by think-aloud interviews, self-reports on a questionnaire, and one-month-delayed recall interviews. Dunlop reported that the three data collection methods yielded very different information regarding learners' attention to feedback. Based on this finding, she made several methodological suggestions on using eye-tracking data for validating language assessment. Firstly, it is important to admit the difficulty of accurately measuring the thinking processes (e.g., attention to reports on test performance) by using the eye-tracking method alone. Secondly, the traditional research methods such as think-aloud, recall interviews and surveys can yield relevant, robust and useful data on cognitive processes if the focus of a study was on how test-takers responded emotionally/attentionally to a diagnostic feedback report on their test performance. Thirdly, as the three research methods were tapping into learners' attention to the feedback report at different points of time, she argued that the different methods were appropriate to address different aspects of the construct of attention.

Schmidt and Pastorino-Campos (**Chapter 8**) presented a literature review of two data collection tools, eye tracking and electroencephalography (EEG), which can be used to tap into test-takers' cognitive processing in language testing. They introduced some key concepts and metrics of eye tracking and EEG in simple language and explained which measures can show which type of cognitive processing as well as their added value to measuring processing load and development of test materials. As shown in the other chapters of this volume, eye tracking has been increasingly used in language test validation, informed methodologically by research in cognitive psychology, psycholinguistics, and language processing and production. According to

Schmidt and Pastorino-Campos, the few eye-tracking studies in language assessment contexts tended to use the eye-movement data to analyse and supplement with other data such as retrospective think-aloud reports from participants. They also pointed out that the studies in assessment contexts have almost exclusively focused on attention-driven processing and suggested that future eye-tracking research in language assessment should focus on the actual underlying cognitive processes and the reasons behind them. They suggested that the co-registration of eye tracking and EEG (i.e., simultaneous recording of brain activity while capturing eye movements) can help better understand what (e.g., the difficulties that test-takers encountered in relation to task features as evidenced in eye-movement data) and why (e.g., the underlying reasons or causes of such difficulties, as evidenced from brain activities recorded by EEG).

Chan (**Chapter 9**) reported a small-scale, exploratory, qualitative study on three university L2 writers' discourse synthesis processes in an integrated reading-into-writing task which requires students to select, organise and connect information from multiple source texts into a new text. The task was part of a post-admission academic literacy test at a British university and is commonly used in higher education contexts. Three sources of data were analysed: students' keystroke logs, retrospective interviews, and the texts they produced in response to the prompts. Overall, the qualitative analysis of the data revealed students' distinct discourse synthesis processes. From methodological perspectives, Chan argued that the qualitative analysis of keystroke logging data, which were often analysed quantitatively in other studies, offered important insights into the cognitive validity of integrated reading-into-writing tasks. Based on the findings of the study, Chan argued that there was clearly a need for explicit teaching and assessment of discourse synthesis processes as well as a rating scale specifically designed to assess relevant processes and skills. She suggested that process-tracking technologies such as keystroke logging may offer an opportunity to incorporate writing processes into assessment criteria that currently focus exclusively on features of the writing output.

Part 2: Using technology to enhance language assessment

Inoue, Nakatsuhara, Berry and Galaczi (**Chapter 10**) reported on a study of the prototype of an IELTS video-call speaking test. Drawing on the notion of context validity in Weir's (2005) socio-cognitive framework, they looked at the administrative features of the test, including examiner behaviours and the effectiveness of examiner training. Advances in video-conferencing technology have made it logistically feasible to have more interactive online communication. During the Covid pandemic, video-conferencing became prevalent in communication. The benefits of video-conferencing

speaking tests are evident, for example, in improving test efficiency and accessibility. Given the different administrative conditions between face-to-face and video-call speaking tests, the study investigated the impact of the key contextual parameters of the video-conferencing mode on both examiners' and test-takers' behaviours. It found that the video-conferencing mode could introduce some changes in examiners' as well as test-takers' behaviours, due to, for example, delayed synchronisation of videos and sound, and less obvious onscreen non-verbal cues for turn-taking or lack of eye contact. These issues could possibly lead to communication breakdowns or increased difficulty in eliciting targeted interactional competence. Examiner training was therefore considered crucial for accommodating such changes in the technology-enhanced speaking test to ensure context validity of the test setting as well as the comparability between video-conferencing and face-to-face delivery of the high-stakes speaking test. However, as the authors rightly pointed out, research on context validity should focus on the comparability in both the opportunities for test-takers to demonstrate their abilities and the rateable sample of language elicited by the tasks, not necessarily in the replication of the testing procedure and timings. They also discussed some broader implications of their findings for other video-conferencing speaking tests and stressed the importance of developing clear and appropriate procedures and guidelines for test administration and examiner training, regardless of what technological innovations are implemented in test delivery.

Link (**Chapter 11**) used Weir's (2005) socio-cognitive framework to guide the development and validation of a computer-based diagnostic assessment tool, called CAFFite, which was conceptualised to support nonnative English-speaking undergraduate and graduate students with their academic writing. The author chose to focus on the scoring validity of the CAFFite algorithms and presented both *a priori* and *a posteriori* evidence. The *a priori* validity evidence was collected at the development stage in which the automated measures used in the algorithms were carefully selected and grounded in two theoretical perspectives, namely Complexity Theory and Systemic Functional Linguistics, to cover the sub-constructs of complexity, accuracy, fluency, and functionality in writing proficiency. The *a posteriori* evidence was about the agreement between CAFFite and human annotators on assessment scores as well as the precision, recall, and F1-score analysis on the writing features identified by CAFFite (on which individualised diagnostic feedback is based). However, Link's evaluation revealed that not all the feedback generated by CAFFite was sufficiently accurate, and the agreement between CAFFite scores and human scores was lower than expected. Link argued for transparency in documenting the progress of developing diagnostic assessment tools even though the results may not be satisfactory or not all evidence was positive, as in the case of CAFFite.

Cotos (**Chapter 12**) integrated the socio-cognitive validity framework (Weir 2005) and the argument-based validity framework (Chapelle 2012, Kane 2013) in validating an automated writing evaluation system, namely the Research Writing Tutor (RWT), which was trained to provide feedback on learners' genre writing competence in academic writing. Based on an evaluation set of 30 Introduction section texts, half written by postgraduate students and the other half collected from research articles published in peer-reviewed journals, the study examined if the RWT classifier could accurately identify the rhetorical traits (i.e., communicative moves and functional steps) in writing based on Swales' (1990) genre theory. Dis/agreement between human annotations and classifications made by RWT were reported as well as the possible underlying reasons for such dis/agreement. The study found that the annotator-classifier agreement was lower than the annotator-annotator agreement and that both of them were lower for step traits than for moves. Cotos attributed the lower annotator-classifier agreement to a limitation in the design of the classifier. That is, while the research article genre is often multi-functional, RWT was only trained to perform mono-label classification. Additionally, she speculated that the potentially different approaches taken by RWT and human annotators in identifying functional cues could lead to discrepancies in classification. It was also reported that RWT and human annotators exhibited similar error patterns. Cotos suggested that functional language (whether present or not) in the texts could be the key sources of mis-classification/annotation of the rhetorical traits. This chapter is a good example of how validity evidence should be examined from different sources and frameworks, and has clearly demonstrated the importance of assessing automated evaluation systems from multiple perspectives as well as the challenges in designing such systems.

Knill and Gales (**Chapter 13**) reported the development of an auto-marker used in a computer-based oral English test called Linguaskill. They described in detail the architecture and training of automatic speech recognition (ASR), a key component that transcribes speech into text in automatic speech evaluation. They also discussed a deep learning-based grader which relies on audio features, text features, and a combination of the two to predict speakers' oral proficiency. According to the authors, the grader draws on an ensemble of multiple Deep Density Networks and the variances across the ensemble are used to indicate the uncertainty of score prediction. Additionally, Knill and Gales explained how the auto-marker deals with abnormal test behaviours, such as a candidate speaking a memorised general answer, meaningless words, or their native language. They concluded the chapter by reporting the evaluation results of ASR and the grader and discussing a hybrid marking approach adopted in Linguaskill that combines automatic marking and human marking to mitigate auto-marker

inaccuracies. This chapter has demonstrated the promises of using machine learning technology in assessing spontaneous second language speech.

Jones (**Chapter 14**) offers a rare insight into the statistical and computational methods that are commonly used or can potentially be used to detect cheating in objectively marked tests. He began by reviewing several major incidents of cheating in large-scale standardised tests and discussing how cheating undermines three aspects of validity in Weir's (2005) socio-cognitive framework. The statistical methods explained in this chapter include person-fit indices, directional copying indices, modelling of response times and score differencing. They are intended to detect candidates' advance knowledge of test items, answer-copying behaviours, and tests sat by imposters. Jones then elaborated on the promise of two computational methods for cheating detection, namely machine learning and cluster analysis, but argued that they are new, unproven, and not as well understood or established as the statistical methods. Finally, Jones reflected on some thorny issues in cheating detection such as ethics of using unexplainable machine learning algorithms, lack of certainty of cheating behaviours, and lack of ready-made software packages for applying the statistical methods.

Enjoy reading the chapters.

In the Epilogue, we will present a broad sketch of what we have learned from these chapters and point to, tentatively, future directions of the role of technology in language test development, delivery and validation research, from the perspectives of the evolving construct of technology-integrated language assessment and the increasing multimodality in human communication.

References

Alderson, J C (2009) Test review: Test of English as a Foreign Language™: Internet-based Test (TOEFL iBT®), *Language Testing* 26 (4), 621–631.

Bernstein, J, Van Moere, A and Cheng, J (2010) Validating automated speaking tests, *Language Testing* 27 (3), 355–377.

Briscoe, T, Medlock, B and Andersen, Ø (2010) *Automated assessment of ESOL free text examinations*, Cambridge: University of Cambridge Computer Laboratory, available online: doi.org/10.48456/tr-790

Burstein, J, Chodorow, M and Leacock, C (2004) Automated essay evaluation: The Criterion Online Writing Service, *AI Magazine* 25 (3), 27–36.

Chalhoub-Deville, M and Deville, C (1999) Computer-adaptive testing in second language contexts, *Annual Review of Applied Linguistics* 19, 273–299.

Chapelle, C A (2012) Validity argument for language assessment: The framework is simple…, *Language Testing* 29 (1), 19–27.

Chapelle, C A and Douglas, D (2006) *Assessing Language Through Computer Technology*, Cambridge: Cambridge University Press.

Chapelle, C A and Voss, E (2016) 20 years of technology and language assessment in language learning & technology, *Language Learning & Technology* 20 (2), 116–128.

Chapelle, C A, Enright, M K and Jamieson, J M (Eds) (2008) *Building a validity argument for the Test of English as a Foreign Language*, New York: Routledge.

Chun, C W (2006) An analysis of a language test for employment: The authenticity of the PhonePass test, *Language Assessment Quarterly* 3 (3), 295–306.

Chun, C W (2008) Comments on 'Evaluation of the usefulness of the Versant for English Test: A response': The author responds, *Language Assessment Quarterly* 5 (2), 168–172.

Clark, T, Spiby, R and Tasviri, R (2021) Crisis, collaboration, recovery: IELTS and COVID-19, *Language Assessment Quarterly* 18 (1), 17–25.

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

Douglas, D (2013) Technology and language testing, in Chapelle, C A (Ed) *The Encyclopedia of Applied Linguistics*, Malden: Wiley-Blackwell.

Enright, M K and Quinlan, T (2010) Complementing human judgment of essays written by English language learners with e-rater® scoring, *Language Testing* 27 (3), 317–334.

ffrench, A, Bridges, G and Beresford-Knox, J (2012) Quality assurance: A Cambridge ESOL system for managing Writing examiners, *Research Notes* 49, 11–17.

Geranpayeh, A (2001) CB BULATS: Examining the reliability of a computer-based test using test-retest method, *Research Notes* 5, 14–16.

Jones, N and Saville, N (2016) *Learning Oriented Assessment: A Systemic Approach*, Studies in Language Testing Volume 45, Cambridge: UCLES/ Cambridge University Press.

Kane, M T (2013) Validating the interpretations and uses of test scores, *Journal of Educational Measurement* 50 (1), 1–73.

Karatay, Y (2022) *Development and validation of spoken dialog system-based oral communication tasks in an ESP context*, Unpublished doctoral dissertation, Iowa State University.

Khabbazbashi, N, Chan, S and Clark, T (2023) Towards the new construct of academic English in the digital age, *ELT Journal* 77 (2), 207–216.

Khabbazbashi, N, Xu, J and Galaczi, E (2021) Opening the black box: Exploring automated speaking evaluation, in Lanteigne, B, Coombe, C and Brown, J D (Eds) *Challenges in Language Testing Around the World*, Singapore: Springer Singapore, 333–343.

LaFlair, G T, Langenfeld, T, Baig, B, Horie, A K, Attali, Y and von Davier, A A (2022) Digital-first assessments: A security framework, *Journal of Computer Assisted Learning* 38 (4), 1,077–1,086.

Leslie, D (2019) *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*, London: The Alan Turing Institute.

Liu, S and Yu, G (2022) L2 learners' engagement with automated feedback: An eye-tracking study, *Language Learning & Technology* 26 (2), 78–105.

Nakatsuhara, F, Inoue, C, Berry, V and Galaczi, E (2017) Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study, *Language Assessment Quarterly* 14 (1), 1–18.

Ockey, G J (2021) An overview of COVID-19's impact on English language university admissions and placement tests, *Language Assessment Quarterly* 18 (1), 1–5.
Ockey, G J and Chukharev-Hudilainen, E (2021) Human versus computer partner in the paired oral discussion test, *Applied Linguistics* 42 (5), 924–944.
Sawaki, Y (2012) Technology in language testing, in Fulcher, G and Davidson, G (Eds) The *Routledge Handbook of Language Testing*, London: Routledge, 426–437.
Schmidgall, J E and Powers, D E (2017) Technology and high-stakes language testing, in Chapelle, C A and Sauro, S (Eds) *The Handbook of Technology and Second Language Teaching and Learning*, Hoboken: Wiley, 317–331.
Suvorov, R and He, S (2022) Visuals in the assessment and testing of second language listening: A methodological synthesis, *International Journal of Listening* 36 (2), 80–99.
Swales, J M (1990) *Genre Analysis: English in Academic and Research Settings*, Cambridge: Cambridge University Press.
Tremblay, A (2011) Proficiency assessment standards in second language acquisition research: "Clozing" the gap, *Studies in Second Language Acquisition* 33 (3), 339–372.
Weir, C J (2005) *Language Testing and Validation: An Evidence-based Approach*, Basingstoke: Palgrave Macmillan.
Xi, X (2008) What and how much evidence do we need? Critical considerations in validating an automated scoring system, in Chapelle, C, Chung, Y R and Xu, J (Eds) *Towards Adaptive CALL: Natural Language Processing for Diagnostic Language Assessment*, Iowa: Iowa State University, 102–114.
Xi, X (2010) Automated scoring and feedback systems: Where are we and where are we heading?, *Language Testing* 27 (3), 291–300.
Xi, X (2022) Validity and the automated scoring of performance tests, in Fulcher, G and Harding, L (Eds) *The Routledge Handbook of Language Testing* (Second edition), London: Routledge, 513–529.
Xi, X, Schmidgall, J and Wang, Y (2016) Chinese users' perceptions of the use of automated scoring for a speaking practice test, in Yu, G and Jin, Y (Eds) *Assessing Chinese Learners of English: Language Constructs, Consequences and Conundrums*, Basingstoke: Palgrave Macmillan, 150–175.
Xu, J (2015) *Predicting ESL learners' oral proficiency by measuring the collocations in their spontaneous speech*, Unpublished doctoral dissertation, Iowa State University.
Yannakoudakis, H, Øistein, E A, Geranpayeh, A, Briscoe, T and Nicholls, D (2018) Developing an automated writing placement system for ESL learners, *Applied Measurement in Education* 31 (3), 251–267.
Yu, G and Zhang, J (2017) Computer-based English language testing in China: Present and future, *Language Assessment Quarterly* 14 (2), 177–188.

Part 1
Using technology to validate language assessment

2 What does the cloze test really test? A cognitive validation of a French cloze test with eye-tracking and interview data

Paula Winke
Michigan State University, USA

Xun Yan
University of Illinois Urbana-Champaign, USA

Shinhye Lee
Michigan State University, USA

Abstract

In a 2011 study that appeared in *Studies in Second Language Acquisition*, Tremblay advocated that a French cloze test she investigated measured French reading comprehension and proficiency. She proposed second language acquisition (SLA) researchers could use the French cloze test to assess the general proficiency of *any French learner* when scores from standardized proficiency tests are lacking, and suggested the test could identify learners along the American Council on the Teaching of Foreign Languages (ACTFL) (2012) proficiency scale. We replicated her study conceptually: we had 22 individuals in a French language programme take the same French cloze test online. In contrast to Tremblay's research study, we tracked the test-takers' eye movements while testing, and we interviewed them afterward. We did this to investigate what the test-takers *do* (via eye-movement records) and what they believed they did (via interviews) during the test. In other words, we used eye tracking and interviews as online research methods during the task, and offline (retrospective) research methods post-task to explore, from two vantage points, the cognitive dimensions of students' experiences in taking the French cloze test. Such research fits within a socio-cognitive framework of test validation (Weir 2005), which is a way to validate test score interpretations. We used the data to investigate whether the cloze test scores represent well the test-takers' true, underlying ability in French, or if the test-takers' individual test-taking processes prevent or prevented, in any way, the scores from being good indications of the

assessed skill. Our data support Tremblay's argument that the test can well discriminate lower-level learners from upper-level ones but, we argue, most tests have such discriminatory power; more difficult is accurately splitting adjacent-performing individuals. Our eye-movement and interview data suggest the test is not appropriate for lower-level learners because the task is too cognitively demanding for them. In this chapter, we use the data to demonstrate that their true French ability was not adequately assessed. We discuss the need for tests that allow test-takers to show what they can do (and *not* what they can't do). Finally, we demonstrate how eye-tracking data, combined with interview data, provide an excellent means to research whether tests allow that.

Introduction

Cloze testing has been claimed to be the most practiced and researched assessment method in language testing (Brown 2013). This may be true when one considers that cloze passages are often employed by teachers as a pedagogical activity, as well as an assessment method, as we will explain in this section. In a cloze test, the test-taker receives a reading passage in which a number of the words are missing. There are multiple ways to 'delete' words or phrases from the passages. For example, the missing words may be evenly spaced (e.g., every seventh or ninth word – a fixed-ration method) or the deletion is linguistically planned (e.g., content words, prepositions or modals are purposefully deleted – a rational method). Likewise, there are multiple ways for test-takers to fill in the blanks. The test-taker may have to think up the missing words on their own, or choose words from a supplied word bank, or answer multiple-choice questions to fill in the gap. In any case, one can think of a cloze test as a *Swiss cheese* reading passage, and the task is to *close up* the passage, that is, make it whole. As described by Taylor (1953:415):

> At the heart of the procedure is a functional unit of measurement tentatively dubbed a "cloze". It is pronounced like the verb "close" and is derived from "closure". The last term is one gestalt psychology applies to the human tendency to complete a familiar but not quite finished pattern to "see" a broken circle as a whole one, for example, by mentally closing up the gaps.

As described by Shohamy (in an interview by Lazaraton 2010), cloze testing has long been seen as a 'magic tool', something easily constructed and graded, cheap and efficient, and 'used to test "everything"' (2010:258). Certainly, cloze passages can be seen as a fun word puzzle, especially when the stakes are low and the task is viewed as a learning exercise or activity in the language classroom. Accordingly, cloze practice books exist that are targeted

for certain age ranges and skill levels. For example, the publisher Scholastic (2012) has a large series of cloze passages that 'help students become active participants in the reading process with engaging, fill-in-the-blank, fiction and non-fiction passages' (2012:4). The passages are labeled with Flesch-Kincaid readability levels (Kincaid, Fishburne, Rogers and Chissom 1975) so that 'teachers choose passages that best meet students' needs' (Scholastic 2012:4). Several websites allow teachers to easily create cloze passages for their students (see a review of them by Ferlazzo (2012)), further demonstrating the popularity of cloze activities in classrooms. In this study, we investigated what learners of French did when they took a cloze test of French. We wanted to see if the passage of the cloze test was appropriate for the students and for measuring their French language ability. We tracked their eye movements while they took the cloze test to monitor their test-taking behaviors, and we interviewed them post-cloze testing to understand how they viewed the French cloze and whether they believed it met their needs.

Literature review

Researchers have been interested in studying cloze processes and constructs because they are popular and an intriguing theoretical puzzle. In 1979, Alderson forewarned that solving the cloze test puzzle was complicated, if not impossible, because text choice, text difficulty level, word deletion method, and what, exactly, gets deleted, all factor into (and change) what is actually assessed. Alderson concluded that cloze testing measures proficiency broadly, but what aspects of proficiency cloze testing in general assesses is not something that can be answered outside of the context of the particular cloze test.

In addition to Alderson's investigation into whether cloze tests measure general proficiency, language testers have investigated the following questions in relation to cloze tests: can cloze measures be good for measuring the proficiency of specific languages, such as Hebrew (Shohamy 1981)? Can cloze tests be used in place of essay tests (Fotos 1991) – that is, can they stand in for writing ability? Do they measure language learners' reading ability (Sadeghi 2014), the reading passage's level of difficulty or readability (Taylor 1953), or something more (Oller 1973)? And are cloze tests a good measure of language, be it general language proficiency (Oller 1973, Tremblay 2011), vocabulary knowledge (Kremmel and Schmitt 2016, Zou 2017), writing ability (Zou 2017), or reading comprehension (McCray and Brunfaut 2018)?

A longstanding and narrower controversy is whether cloze test-takers focus only on phrases, sentences, or read the entire text (Bachman 1985, Brown 2013, Yamashita 2003). Researchers have also investigated whether cloze tests measure lexico-grammatical knowledge only or global reading

comprehension (Brown, Trace, Janssen and Kozhevnikova 2016). But the results as to what, exactly, test-takers focus on during cloze testing are mixed. For example, researchers observed that when cloze sentences were scrambled, test-taker performance remained the same (e.g., Markham 1985). However, others found that test-taker performance *dropped* after sentence scrambling; they therefore argued that cloze tests measure a learner's ability to utilize information across sentence boundaries (e.g., Chavez-Oller, Chihara, Weaver and Oller 1985, McKenna and Layton 1990). Others have investigated whether cloze test constructs depend on the closeness of the semantic relationship between the deleted words and their surrounding text (e.g., Bachman 1985, Kobayashi 2002, Trace, Brown, Janssen and Kozhevnikova 2017). The researchers speculated that the deleted words whose meaning could be inferred within the immediate sentential context might only measure local lexico-grammatical knowledge, but deleted words whose meaning must be inferred from across sentences would require global reading comprehension. The newer view is that a cloze test can measure both local lexico-grammatical knowledge and global reading comprehension, but it depends on the context and the examinees (McCray and Brunfaut 2018, Yamashita 2003).

Within the ocean of research on cloze testing, one study stood out to us when it was published because it seemed to broaden the accepted use of cloze testing, rather than restrict or hedge its 'magical' application: Tremblay (2011) proposed that a single cloze test can be a reliable and valid measure of global language proficiency for second language (L2) learners at all levels of proficiency. More specifically, the author suggested that a French cloze test that she and a colleague (Tremblay and Garrison 2010) developed could be used to place students on the ACTFL proficiency scale (2012), from Novice to Advanced. The author did not speculate on whether that would be on the ACTFL proficiency scale for reading, writing, or listening – she did note (2011:363) that the cloze test is limited in assessing oral proficiency, so one could reasonably assume that the proficiency level designation would be for reading (or perhaps writing) proficiency. Tremblay wrote that her cloze test could 'not only enhance the interpretation of experimental results but also decrease the disparity between the proficiency assessment methods used in SLA research and thus facilitate comparisons between studies' (2011:364).

Tremblay (2011) administered her and Garrison's (2010) French cloze test to 169 university French language learners. Overall, the test had a high reliability estimate (KR-20 = .92) and a high ability to differentiate the top third of test-takers from the bottom third (with a mean discrimination index of .41). Tremblay used a regression analysis to demonstrate that a composite language proficiency estimate (based on a number of variables such as age of first exposure, years of instruction, and self-rated proficiency)

predicted cloze test outcomes (β = .73). Based on these findings, Tremblay stated that the cloze test met both testing and SLA research standards, and could be used to measure the general language proficiency of learners at all levels for both research and instructional purposes. Since publication, Tremblay's article has been cited by the researchers of at least 29 studies, with most of them citing it as justification for using a single cloze test to measure foreign or second language proficiency. Thus, the impact of Tremblay's work on research in the field of SLA is large, and it is time for a second opinion look at her claims that (a) the French cloze test measures all levels of French proficiency, and (b) the test is appropriate for any college learner of French.

Tremblay's strong arguments and the impact the article has had on the measurement of proficiency in SLA research prompted us to conduct a *conceptual replication* (Polio 2012:51) of her study. Because we were interested in calling into question Tremblay's findings (a common motivation for replication research; see Polio (2012:83, citing Santos 1989)), we wanted to replicate her quantitative findings but triangulate them with eye-tracking metrics and participant interviews to understand better not just how test-takers scored, but also what processes they used to respond to the items and how they perceived the test itself.

We added eye tracking and interviews because it is important to understand test-takers' *response processes*. Indeed, response processes are a source of validity evidence called for by the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME] and Joint Committee on Standards for Educational and Psychological Testing 2014). The idea, as described by Messick (1990), is to 'directly probe the ways in which individuals cope with the items or tasks, in an effort to illuminate the processes underlying item response and task performance' (1990:1,488). Weir (2005) conceptualized Messick's notion of researching test-takers' item-response processes to understand whether the test score interpretations are valid. This process of investigating what test-takers think and do while taking the test is a process of establishing the *cognitive validity* of the test score interpretations. Weir (2005:63) wrote that if the cognitive processing involved in determining test answers (on a reading test) bears little resemblance to the way individuals process texts for information in real life, then the assessment 'may be considered deficient in terms of theory-based validity'.

The cognitive validity question is whether a test elicits the mental processes that the individuals would employ in real-world conditions that are not part of any test. Cognitive validity is part of construct validity, and 'refers to the relative dependence of task responses on the processes, strategies, and knowledge (including metacognitive or self-knowledge) that are implicated in

task performance' (Messick 1995:742). Investigating the cognitive processes involved in responding to items on a test can help researchers uncover sources of construct invalidity. The construct, according to Messick (1995), may be under-represented (the test fails to include necessary dimensions of the construct), or the targeted construct may actually just be one part of what is assessed. In the latter case, the threat to validity is construct-irrelevant variance, which is when other, distinct constructs (or things like guessing) muddy the water and make scores dependent on something other than the construct intended to be assessed. Cognitive validity is one way to approach the design of a test validity argument, and one can think of test validation as important before the test scores are put to use. But continuing to investigate a test's validity after the test is operational is important as well, because test score uses change, and tests themselves must change (be revised, adapted, or have items that get refreshed) so that they remain current, authentic, and do not produce results that are skewed due to out of date material. Only by understanding how test-takers cope with the test tasks can researchers know and verify what the test measures, as scores alone cannot provide that information. Taking a cognitive approach to test validation (as outlined in Khalifa and Weir 2009), in this study we investigate the mental processes that cloze test-takers use to fill in the gaps in a cloze reading test passage. We assume, as cloze test theory posits, that test-takers read the text, and that they fill in the blanks by reading the words and phrases that occur before and after the blanks.

Thus, in this conceptual replication, we addressed the following two research questions (RQs). The first RQ examines whether test-takers' eye movements on the cloze test resemble a largely linear, sequential movement pattern of reading. The second RQ addresses the test-takers' perceptions of the cloze test. We ask this question to elicit their perceptions about the difficulty and appropriateness of the tasks to measure their language proficiency and to obtain a deeper understanding of their eye-movement pattern.

RQ1: What are test-takers' response processes when they take the Tremblay and Garrison (2010) cloze test? For example, do the test-takers show evidence of reading, and if yes, what type of reading (local, lexico-grammatical knowledge or global reading comprehension)?

RQ2: How do the test-takers perceive the test? Specifically, do they perceive the test as a reading comprehension test that assesses their French reading ability (or perhaps as a writing test), or do they perceive the test as a French grammar and vocabulary test, or as a general test of French language proficiency?

Methods

Participants

The participants comprised 22 individuals (10 males, 12 females). Like Tremblay (2011), we recruited the participants through a French department at a large mid-western university in the US. The sample size of this study might be considered small compared to the English as a Second Language/English as a Foreign Language (ESL/EFL) contexts; however, obtaining a large sample of participants is less feasible given the much smaller learner population in foreign language programs in US universities. To better represent the target learner population, we purposefully recruited four students each from first and fourth year French classes, and five students each from second and third year classes (with the classes belonging to a 4-year curriculum at the university that is designed to take students from Novice to Advanced-low proficiency on the ACTFL (2012) proficiency scale). At this particular university, as is common in many US universities, first year classes are called 100-level classes; second year classes are 200-level; third year are 300-level; and fourth year are 400 level. Among the students in this study, 18 (seven males, 11 females) were undergraduate students, aged 18 to 23. We also recruited two graduate students and two instructors of French to mirror the participant make-up of Tremblay's study. One graduate student was enrolled in the French first year (100-level) course, and the other was a native speaker of French; both were over the age of 25. Of the two French instructors, one was a native speaker. All other participants were native speakers of English. A full list of the participants, along with summaries of their demographic information (which we will explain below in the 'Results' section), are in Appendix 1.

Materials

Computerized cloze test

We created an online version of Tremblay and Garrison's (2010) paper-based cloze test in order to put it on an eye-tracking computer (see Appendix 1, Tremblay 2011). The cloze test features a 314-word non-technical article about global warming, with 45 blanks following the rational deletion method. We added an onscreen feature that allowed participants to copy and paste French accents (see Figure 1). In Tremblay's study, the cloze test was on paper, and the entire passage, we assume, was on one page. Similarly, in this study, the cloze test was on one single web page (see Figure 1, and see Winke (2024), for anonymized video recordings of Participants 2 and 22 taking the test online). Individuals could scroll the web page during the test. Admittedly, the paper-based version of the cloze test and the computer-based version of

the cloze test are not the same, but we argue that, in this case, test delivery mode should not make too much of a difference, as college students are used to reading online, and are used to fill-in-the-blank activities online. (Although certainly, with children or within testing contexts of higher stakes, test mode matters.) Putting the cloze test online was a methodological decision that we made in part so that we could easily track the participants' eyes while they read. Also, short and quick tests of proficiency today most certainly need to be online for practical reasons and for ease in delivery and scoring.

Post-experiment questionnaire and interviews

We created a post-experiment questionnaire, via Tobii Studio (www.tobii.com), to gather information about participants and their language background (e.g., age, gender, year in college, whether or not they had studied abroad). We also asked them, in one-on-one semi-structured interviews, to self-assess their French reading ability, their comprehension of the cloze passage, the difficulty of the passage, and to tell us the main idea of the passage. We asked them their perceptions of the test, and how they responded to the test items. A summary of the questionnaire and interview responses are in Appendix 1, and a complete list of interview questions can be found in Appendix 2.

Procedure

Participants met individually with a research assistant in the Tobii TX300 (www.tobii.com) eye-tracking lab. The Tobii we used for this research recorded at 300 hertz (Hz) and used binocular eye tracking (both eyes were tracked). Prior to test taking and recording, each participant's eyes were calibrated to the eye-tracking camera, as is standard with the Tobii TX300 system.

As in Tremblay's study, each participant took the cloze test at their own pace (the test was not timed). We used Tremblay's written instructions. After the test session, the participant responded to the questionnaire and participated in the semi-structured interview, which was audio-recorded. We did not show the participant their recorded eye-movement video. In other words, we did not use the eye-movement recording video as a stimulus because we felt the stimulus would have been too strong in this case, and may have potentially distracted the participants from thinking about what they were thinking at the time of testing. (Note, however, that other researchers have taken the view that showing participants their eye movements is facilitative, e.g., Brunfaut and McCray 2015.) Thus, we only showed the participants the blank test on screen so that they could look at the blank test during the interview. The time spent on the entire procedure varied, ranging from 20 minutes to one and a half hours (see Appendix 1 for each person's exact time). All participants were paid $20.

Figure 1 A computerized version of the French cloze test from Tremblay (2011)

Personal Info

First Name _____

Last Name _____

Email _____

Character Map

À	à	Â	â	Æ	æ	Ç
ç	È	è	É	é	Ê	ê
Ë	ë	Î	î	Ï	ï	Ô
ô	Œ	œ	Ù	ù	Û	û
Ü	ü	«	»	€	₣	

Le taux de CO2 dans l'atmosphère augmente plus vite que prévu

La croissance économique mondiale _____ provoqué un accroissement de _____ teneur en dioxyde de _____ (CO2) dans l'atmosphère beaucoup _____ rapidement que prévu, selon une étude _____ lundi dans les comptes rendus de l'Académie _____ des sciences des États-Unis. Cette étude _____ que la concentration des émissions _____ gaz car-bonique dans l'atmosphère a_____de 35% en 2006, entre le début _____ années 1990 et les _____ 2000–2006, passant de 7 à 10 milliards de tonnes_____ an, alors que le protocole de Kyoto prévoyait _____ en 2012, ces émissions responsables _____réchauffement climatique devaient _____ baissé de 5% par_____ à 1990. "Les améliorations dans l'intensité carbonique de l'économie _____stagnent depuis 2000, après trente _____ de progrès, ce qui a provoqué cette _____inattendue de la _____ concentration de CO2 _____ l'atmosphère", indique dans _____ communiqué le British Antarctic Survey, _____ a participé à cette étude. _____ les chercheurs, les carburants polluants _____ responsables de 17% de cette augmentation, _____ que les 18% restant sont _____ à un déclin de la capacité des "puits" naturels comme _____ forêts ou les océans _____ absorber le gaz carbonique. "_____ y a cinquante ans, pour chaque tonne de CO2 émise, 600 kg _____ absorbés par les puits naturels. _____ 2006, seulement 550 kg par tonne ont été _____ , et cette quantité continue à baisser", explique _____ auteur principal de l'étude, Pep Canadell, du Global Carbon Project. "La baisse de l'efficacité_____ puits mondiaux laisse _____ que la stabilisation de cette _____ sera encore plus _____ à obtenir que ce que l'on pensait jusqu'à _____ ", indique pour sa _____ le British Antarctic Survey. Ces _____ obligent à une révision à la hausse _____ prévisions du Groupe intergouvernemental d'experts _____ l'évolution du climat qui, dans son _____ de février, tablait sur l'augmentation de la température _____ de la terre de 1,8°C _____ 4°C _____ l'horizon 2100.

Submit

Analysis

Item and test scores

We followed Tremblay (2011) in scoring the test by assigning one point to correct responses and a zero point for incorrect responses. Our scoring was based on Tremblay's list of acceptable answers. We calculated the item facility (IF) and item discrimination (ID) indices for all cloze items (for an overview of these indices, see jalt.org/test/bro_17.htm) as well as the overall test IF, ID, and reliability indices.

Questionnaire and interview data

We transcribed the interviews. Then, we coded the interview and questionnaire responses in ways that allowed us to sort or average test-takers' responses to each question. For example, on the question 'Did you think the test was appropriate for you?', we coded the individual responses as 'yes', 'no', or 'unsure'. For the question 'Did you get the main idea of the passage?', we coded the responses as 'yes' when they showed participants' correct understanding of the test passage and 'no' when they were off-topic (e.g., Participant 22's response: 'I think it is about American history'). We also asked, 'What language skill(s) do you think was measured by the test?' We coded participants' responses in short phrases (e.g., science knowledge, Participant 14). Each participant's coded responses can be found in Appendix 1.

Eye-tracking metrics

Prior to analyzing the eye-tracking data using IBM SPSS version 24, we designated certain parts of the cloze test as *areas of interest* (AOIs). The Tobii TX300 collects eye-movement records within an AOI and then calculates raw metrics per person, which indicate how much time (in seconds) or how many times (in frequency counts) the participant looked at the information within the AOI. We created AOIs around the individual blanks as well as the words or phrases before and after each blank, amounting to a total of 138 AOIs (46 cloze items by three AOIs) for analysis. We used our own judgment to decide how many words before or after each blank defined an AOI. As a general rule, we kept semantic units intact and incorporated them into the AOI, such as prepositional phrases or phrasal verbs, or short, highly frequent phrases. The variation in AOI length is messy and contributes to somewhat messy data, but because language itself is messy and our study is exploratory, we do not believe the inherent messiness affected the results in a significant way. We had to make these AOIs semi-dynamic because we allowed for scrolling;

that is, after the eye movements were recorded, we opened each person's eye-tracking record and manually moved the overlay of the AOI template (the keyframe) to keep the AOIs attached to the right areas (the text) on the keyframe. (How to do this is further explained in the *Tobii Studio User's Manual* (2016).) Figure 2 is a screenshot of the AOIs. The AOI outlines were not visible to the participants.

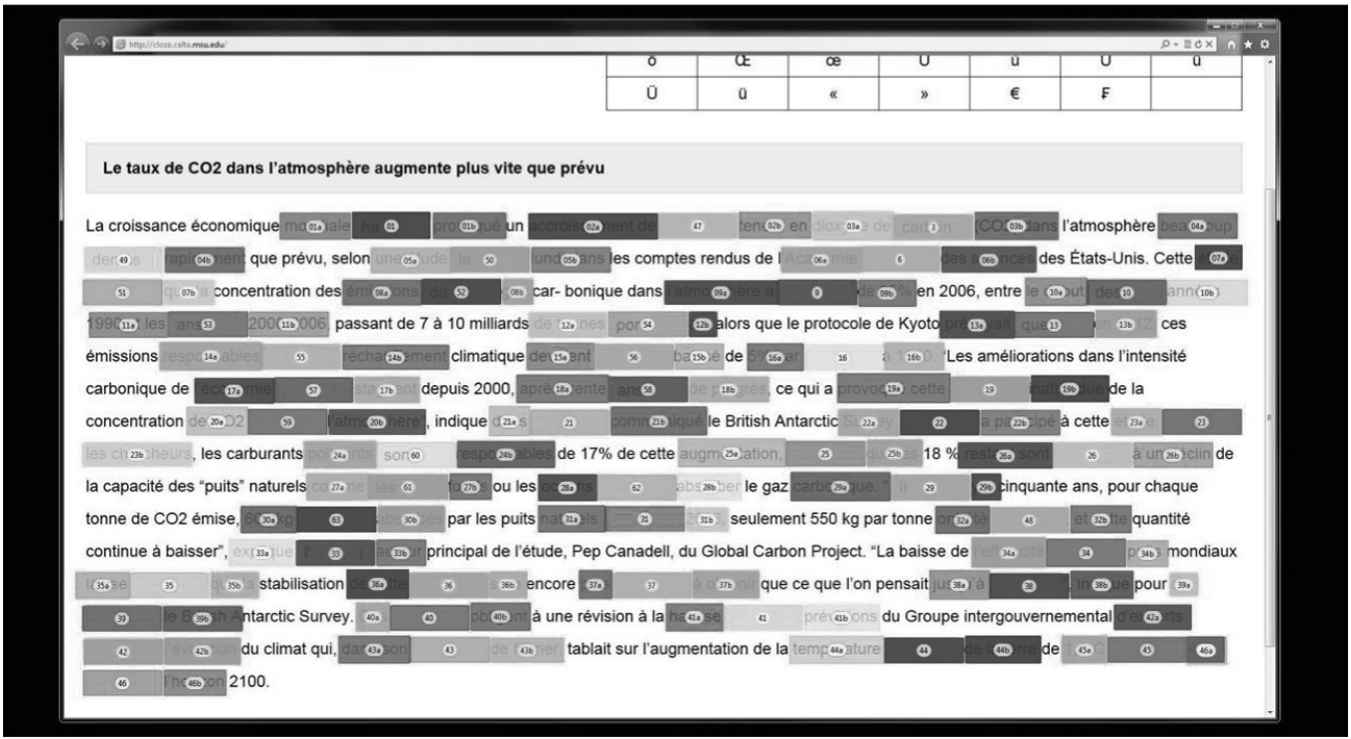
Within each AOI we collected two different types of eye-movement data: a) the *time to first fixation* (TFF), which measures (in seconds) how long it takes (after the start of recording) before a participant fixates their eyes on an AOI; and b) the *total fixation duration*, which measures (in seconds) the sum of the duration of all eye fixations within an AOI. We hypothesized that the TFF can evidence whether cloze items measure local, lexico-grammatical knowledge only or global reading comprehension. Specifically, if the test prompts the latter, we would expect participants to process the cloze test items sequentially, thus one would see a strong positive correlation between TFF and AOI number. In contrast, if cloze testing mostly measures lexico-grammatical knowledge, the relationship could be weak or non-significant, as the participant need not read the text sequentially (see, e.g., Reichle, Pollatsek, Fisher and Rayner 1998) but can instead jump back and forth (i.e., local-level processing). With the total fixation duration, we measured the amount of time participants spent looking at the individual cloze items, along with the words/phrases before and after the blank. We hypothesized a negative relationship between the items' facility levels (IF indices) and the amount of time test-takers spent reading the words around the blanks: more difficult items (which have lower IFs) would attract longer visual attention than easier items (which have higher IFs).

Results

Summary of test-takers' scores

The average test score was 16.45 (SD = 11.46; Range = 0–45) out of a maximum of 46. The standard deviation of the cloze test scores was large, as expected, because the students themselves ranged widely in terms of their French ability: we pre-selected students who ranged from near beginners in French, to advanced speakers and users of the French language. Overall, the participants' level of French study corresponded with the test results (as in Tremblay's study) because she too sampled from learners who ranged widely in terms of their French ability. In our study, the French instructors (Participants 2 and 3) and the native speaker (Participant 1) scored the highest. French course level (e.g., first year, second year) correlated with test results (r = .85), as did study abroad background (r = .70), self-assessed French reading proficiency (r = .80), and self-assessed comprehensibility of

Figure 2 A computerized version of the French cloze test from Tremblay (2011), with the AOIs indicated



the cloze passage (r = .70). Overall, the test had a reliability coefficient of .990 (KR-20). The ID indices (calculated as the difference in IF between the upper- and lower-third test-takers) ranged from .05 to .91. The mean ID was .37 (SD = .25), which was similar to that in Tremblay (2011:354) (ID = .41).

The test times varied drastically among participants, from three minutes to over an hour. For example, Participants 4 and 5 each received a final score of 25, but Participant 4 finished in 13 minutes, while Participant 5 finished in one hour and two minutes (see Appendix 1). In Tremblay (2011), the test-takers 'usually required between 15 and 35 minutes to complete the test' (2011:352), which prompted her to argue for the efficiency and practicality of the test (2011:360). However, in our study, the wide range of test times, not just the averages, suggests otherwise.

Eye-tracking metrics

To investigate test-takers' overall response processes, we first correlated the participants' eye fixation durations on the AOIs before and after the cloze test blanks with the items' IFs. Instead of performing these correlations for each individual (amounting to a total of 44 correlations: two correlations per test-taker), we divided the test-takers into two groups according to their test scores: upper-level performers (N = 11) and lower-level performers (N = 11). With two groups and three sets of AOIs (the blanks and the words/phrases before and after them), we computed six correlations (Pearson's r). We predicted a negative correlation between IF and total fixation time. Interestingly, we found significantly negative correlations for the higher-scoring participants (with significance set at .05) between IF and total fixation durations on words before and after the blanks ($r_{before} = -.352$, $r_{after} = -.303$), but not on the blanks themselves ($r_{blank} = -.210$). However, these relationships weakened for the lower-scoring participants ($r_{before} = -.193$, $r_{blank} = -.174$, $r_{after} = -.238$) (see Table 1).

This finding suggests that the relationship between visual attention and item difficulty is moderated by learners' language proficiency level. That is, higher-level performers attended to the contextual information longer around more difficult items than easier items. Lower-level performers' attention, on the other hand, had a weak or non-significant relationship with the items' facilities.

To explore whether the individual participants processed the cloze passage in a sequential manner (which we argue can evidence cloze tests as a measure of global reading comprehension), we ran correlations between individual participants' TFF on the AOIs and the AOIs' numbers (i.e., 1–46). All participants except one (Participant 10, an outlier in regard to the eye-tracking metrics) suggested that there is a significant positive correlation between these values (see Table 2). However, the relationships

Table 1 Correlation (Pearson's r) between the AOI's total fixation duration and item facility

Group	AOI location		
	Before blanks	Blanks	After blanks
Upper-level performers (N = 11)	–.352* (p = .03)	–.210 (p = .16)	–.303* (p =. 04)
Lower-level performers (N = 11)	–.193 (p = .20)	–.174 (p = .25)	–.238 (p = .11)

**p value is significant (with p less than .05)*

Table 2 Correlations between the participant's TFF and the AOI order

ID	Test score	Test time	Correlation between TFF fixation and item order (1–46)		
			TFF and order of AOIs before blanks	TFF and order of AOIs that are blanks	TFF and order of AOIs after blanks
1	45	15:21	.906** (40)	.754** (43)	.812** (41)
2	42	7:43	.965** (41)	.875** (40)	.962** (38)
3	33	15:07	.828** (44)	.559** (45)	.536** (45)
4	25	13:03	.973** (34)	.856** (34)	.929** (35)
5	25	1:02:01	.844** (46)	.874** (46)	.831** (46)
6	21	23:56	.830** (42)	.753** (43)	.777** (41)
7	19	12:53	.515** (45)	.286 (43)	.503** (45)
8	18	26:18	.696** (46)	.631** (46)	.540** (46)
9	18	15:00	.674** (39)	.709** (41)	.687** (39)
10	18	21:49	−.437 (3)	.954 (3)	.893* (6)
11	17	45:29	.628** (46)	.835** (46)	.576** (46)
12	16	15:53	.692* (11)	.835** (15)	.576** (15)
13	15	44:24	.644** (39)	.635** (39)	.597** (41)
14	14	23:31	.935** (45)	.936** (46)	.838** (44)
15	12	11:45	.948** (44)	.964** (42)	.924** (43)
16	9	44:11	.516** (46)	.469** (46)	.458** (46)
17	8	9:52	.631** (33)	.800** (34)	.632** (34)
18	8	15:15	.945** (45)	.964** (46)	.739** (46)
19	3	8:25	.805** (37)	.816** (35)	.852** (38)
20	3	10:17	.937** (40)	.838** (38)	.936** (44)
21	2	13:37	.938** (43)	.927** (42)	.968** (42)
22	0	3:41	.869** (29)	.729** (23)	.785** (21)

Note. Number of items included in the final analysis (that is, the number of items fixated on) is indicated in parentheses. Numbers vary because not all participants read or looked at all items.
** Correlations significant at .05; ** Correlations significant at .01*

were on average stronger for higher scorers (those scoring 21 or above, e.g., Participants 1 and 2) than lower-level test-takers (e.g., Participants 7 and 16). This sequential pattern of reading suggests that higher-proficiency learners tended to demonstrate a clear pattern of global reading (i.e., they read the passage sequentially and extracted both intra- and intersentential

information to complete the cloze items), which included a certain amount of skipping, regressing, and other natural reading behaviors. In contrast, lower-proficiency learners tended to not read the entire passage, skip complete sections entirely, and instead used guessing strategies frequently (i.e., looking back and forth between blanks, not written words, to seek items that could be completed without having to understand the whole passage). This explanation is also supported by the number of AOIs individual participants skipped (an AOI is 'skipped' when the eye tracker registers no fixations within it; see Table 2). For example, Participant 22, who scored 0 on the cloze test, skipped 17 of the 46 AOIs before the cloze blanks. This pattern of high-level skipping indicates that the participant did not look at or read a large proportion of the cloze test passage.

Qualitative (interview) summary

In the supplementary file, we summarized the participants' responses to the post-experiment questionnaire and interviews. Higher-level performers (Participants 1, 2 and 3) reported complete comprehension of the text, whereas other participants' self-assessments of text comprehension tended to range from medium to high (5 to 7). Likewise, the majority of the participants (except Participant 22) were able to articulate the main idea of the passage (carbon dioxide emissions).

Although 54% of the participants (n = 12) considered the cloze test as an appropriate measure of their language ability, most (including all the 100-level participants and the two French instructors) perceived the passage as too difficult, especially for lower-level students. For example, Participant 17 mentioned the following (comments are unedited to maintain authenticity):

> I feel like it's [the test is] for … France people. It's so hard, especially for people like me, as far as studying for a year, and then you show up with a test like that. If that's my class, I would fail it for sure. I would fail so miserably.
>
> (Participant 17, male, senior, French 100-level, test score 8)

Higher-scoring participants perceived the test as suitable for advanced-level French learners. For example, Participant 1 (a native speaker of French) stated:

> It [the passage] is talking about a subject that's not very easy, so it's not like they're talking about kids playing or something like that. I mean, it's a subject that you need to … know a little bit in French to be able to … tackle in this text. So it's a good test.
>
> (Participant 1, male, native speaker, test score 45)

Interestingly, when asked about the skills measured by the cloze test, 81% of the participants (n = 18) considered the cloze test as a measure of lexico-grammatical knowledge. Three participants mentioned reading, and two thought it tested memory (Participant 18) or guessing skills (Participant 2). Among them, 12 participants (54%) reported relying on sentential context to derive answers. For example, Participant 6 described her process of extracting grammatical information from the surrounding words to infer the correct part of speech for the blank:

> [I was] most of the time looking at the verb, or the word in front of the blank and the word behind the blank … less so in the context of the sentence, I guess, and more in the sentence structure, because if there was a past participle after the blank, obviously, the blank was going to be *avoir* or *être*, conjugated in the past tense. And then if there was a *de* in front, and then, like a noun, then maybe, the space was *la*, if the noun was feminine. So not really in the context of the sentence, though.
>
> (Participant 6, female, junior, French 400-level, test score 21)

However, higher- and lower-scoring participants seemed to differ in their test-taking behaviors. Three high scorers (including Participant 1, the native speaker) stated that they read the passage or parts of the passage multiple times to get a sense of the context. On the other hand, five low scorers reported only selectively reading the text or scanning the text for similar words or phrases elsewhere to help fill in the blanks. For example, Participant 16 stated:

> Umm, a lot of times, if I couldn't think of what it [the missing word] was right away, if I didn't think I knew what it was, I'd look throughout the rest of the text to see if there was anything similar to it … or something I could've used from other places throughout the text.
>
> (Participant 16, male, senior, French 200-level, test score 9)

Discussion

How our results compared to Tremblay's (2011)

In this conceptual replication of Tremblay (2011), we re-examined the reliability and validity of Tremblay and Garrison's (2010) cloze test on a similar (but smaller) sample of French learners. While our quantitative results aligned in almost every way with Tremblay's, our additional qualitative data led us to very different conclusions. Tremblay noted the cloze test is a good measure of global proficiency, but our eye-tracking results indicated a wide dispersal of test-taking behaviors, demonstrating that scores appear to represent either linguistic knowledge and skills (for the

higher-level test-takers), or the ability to manipulate linguistic information with or without comprehension or to guess correctly (for the lower-level test-takers). This suggests that global test score interpretation is problematic. On another note, our data partially supported Tremblay's claim that the test can help select a homogeneous group of language learners for research purposes, but again, our qualitative data forces us to add the caveat that this type of testing with this specific test is only suitable for higher-proficiency learners.

More importantly, our results led us to reject Tremblay's claim that this particular cloze test can be used to determine or diagnose students' proficiency levels in reference to the ACTFL scale (Novice through Advanced). We reject this because such score band assignment based on this test would be unfair for lower-level learners. We explain our interpretation of the study's data more in the following section.

Why the cloze test cannot fairly measure Novice or Intermediate proficiency

Admittedly, cloze tests can reliably differentiate language learners across proficiency levels in a quick and inexpensive fashion (e.g., Brown 2013, Oller 1973). However, from an assessment or measurement perspective, discriminating power is far from enough to justify all purposes of test use (AERA et al 2014). The test must also be fair and the scores useful. Tremblay noted that the test scores could reliably separate test-takers into four distinct groups and that these levels could be bookmarked 'to scale descriptors such as Novice-high, Intermediate-low, Intermediate-high, and Advanced' on the ACTFL proficiency scale (Tremblay 2011:361). We disagree with this as an overgeneralization. Our eye-tracking metrics indicate that the lowest-level test-takers could not read or comprehend the passage and, instead, relied on guessing strategies to fill in the blanks (such as Participant 16). In the interview data, they expressed being unable to demonstrate their actual language ability on such a difficult test; some even developed negative feelings for not being able to understand the bulk of the text. Tremblay (2011) recognized this problem and pondered 'whether administering the present cloze test to such L2 learners is ethical, as the participants will not gain anything from it, and for that matter, nor will the researchers' (2011:363). Thus, based on this research, we believe this cloze test really tests upper-level French reading and grammar knowledge, and a low score can only indicate that the test-taker does not have upper-level French reading and grammar. What level of French reading and grammar they are at is not assessed. In fact, Tremblay (2011:345) gave the same opinion:

> As Messick (1989) suggested, validity is not a property of a test but rather of the inferences made on the basis of the test (for a discussion,

> see also Chapelle 1999). For such inferences to be accurate, the cloze test must be neither too easy nor too difficult for the targeted L2 learners; otherwise, the test may not reveal much about these learners' proficiency other than whether it meets a particular level.

In language assessments, learners should receive texts that are appropriate for their level to demonstrate what they really can do (what they can read); otherwise, the assessment would not be fair. Tests are fair when the test-takers can 'demonstrate their standing on the construct(s) the test is intended to measure' (AERA et al 2014:51). It may be unethical to present this test to lower-level learners because these learners may be misled to assume that they were *supposed* to be able to understand the passage that was way above their (expected) achievement level. Applied linguists should not give tests that mislead language learners into thinking they should know more than they should, even when the testing is for research purposes only.

Another conceptualization of test fairness, which is part of test validity, states that tests need to have '*comparable validity* for all test-takers and test-taker groups' (Willingham and Cole 1997:6–7; emphasis added). Test results first need a small overall statistical error (as is found in both Tremblay (2011) and in our current conceptual replication study), but the error should also be comparable across test-takers and test-taker groups. This assumption largely holds true if all examinees apply their actual ability, knowledge or skills to respond to test items. However, guessing leads to test measurement error, 'contaminating' test-takers' scores and muddling construct representation (Messick 1996:4), and there is evidence in our data that the lower-level test-takers guessed more than upper-level ones did. Thus, it could be seen that the lower-level students' scores were more prone to measurement error than the upper-level students' scores were, suggesting the test does not have comparable validity across groups.

Comparing the cloze passage on carbon dioxide emissions to ACTFL's (2012) reading scale and ACTFL's online description of an Advanced text, we believe the cloze passage is Advanced in terms of text difficulty. At the Advanced level, texts are authentic and describe real-world topics of interest; readers with Advanced-level proficiency can read subject matter that is concrete and understand most vocabulary in Advanced texts, except for occasional words and phrases for which the readers may need to use contextual clues to derive meaning (ACTFL 2012:22). Thus, this cloze passage should not be presented to those who are far below the Advanced level in reading. We believe our research shows what happens when Advanced-level texts are presented to Novice-level learners: the Novice students cannot do the task, and thus their actual performance ability is not captured by the test. Our research underscores the need to carefully choose the passage for cloze testing (as recommended by Trace et al 2017) so that the

passage difficulty approximately matches the test-takers' language ability. This match harks back to the notion that cloze tests are tests of readability (Taylor 1953). In light of our findings, this original consideration seems most right. A cloze passage at the Advanced level can best measure a test-taker's reading comprehension at the Advanced level. Scoring too high or too low would only indicate that the test-taker is not at the Advanced level. Applied linguistics researchers need not use Flesch-Kincaid readability levels to index their cloze passages, as Scholastic (2012) does (although they certainly can), but can (and perhaps must) benchmark their cloze passages to standardized proficiency descriptions that applied linguists use more commonly, such as the ACTFL reading levels. Such benchmarks would help SLA researchers choose cloze tests that best match their language teaching and assessment needs.

Conclusion

The question we presented in our title was 'What does the cloze test really test?' We feel our eye-tracking and interview data provide further evidence that prior researchers have been right. Cloze tests measure reading ability (Kincaid et al 1975, Taylor 1953), and because certain levels of vocabulary knowledge and grammar are needed to read certain texts (which are at certain difficulty levels), cloze tests also can be seen as indirect measures of vocabulary level (Kremmel and Schmitt 2016, Zou 2017) and, perhaps, grammatical knowledge. The students' interview data contributed to such interpretations. As readability tests, cloze tests may be best at answering *only* whether the person being tested can read the text (yes, the person can read this text, and possibly texts like it at this difficulty level, or no, the person cannot read this text, nor other texts like it at this difficulty level). With a cloze test, one can answer the question of whether the person being tested has the functional language ability to read and comprehend the text being presented. If the text is at the Advanced level, a cloze test can really only test whether the person being tested can read and comprehend that Advanced-level text, which may indicate his or her ability to read at the Advanced level in general (which could be seen as an Advanced proficiency level check, as alluded to in Alderson (1979)). But the cloze test really can't do more than that, especially because it is a cloze test with a single passage. Based on the students' interviews, we can see (as we expected) that the cloze test in this study did not tap into the students' real-life writing processes: thus, using a cloze test as an indication of writing ability would be a risky endeavor, and counter to the cognitive validity evidence collected in this study.

Tremblay's (2011) findings and recommendations may have been seen by SLA researchers as a welcome solution to a longstanding problem: there is a lack of short, inexpensive assessments of language proficiency for

SLA research that are both economical and reliable. Authors of research methodology guidelines have mandated that researchers include external measures of language proficiency in their studies and no longer simply report the participants' length of language learning (one year, two years) or course level. As Tremblay (2011) explained, time learning the language can vary qualitatively, and language classes and learners are not homogeneous, thus external measures of proficiency are needed for SLA research. However, we warn that a one-size-fits-all approach to language proficiency assessment might not be ideal, especially when the assessment relies on a single cloze passage. Admittedly, the problem with the Tremblay study is not necessarily that her cloze test measures something other than what she claims it did, but that she advocates using a single test passage for a wide spectrum of proficiency levels. In practice, such a passage might be too difficult for some learners, thus providing limited diagnostic information about their language performance; alternatively, the passage might be too easy for others, which presents a problematic ceiling effect, especially if the assessment purpose is to check for proficiency variation. A promising solution is to create computer-adaptive cloze tests with a series of passages that partly overlap but gradually increase in text and item difficulty. This type of work is ongoing at the Georgetown University Assessment and Evaluation Language Resource Center (AELRC; aelrc.georgetown.edu/current-projects), where researchers are developing computer-adaptive C-tests (a similar general proficiency measure to cloze tests) for measuring proficiency in Arabic, Japanese, Russian, Portuguese, Turkish, and soon for Bangla and Korean. Similar to Tremblay (2011), AELRC aims to present a measure of language proficiency for L2 research purposes. Another similar project is being undertaken by researchers from the European Association for Language Testing and Assessment (EALTA): this group is creating computer-adaptive C-tests with multiple, overlapping texts of different levels of difficulty for English, Finnish, French, German, Italian, and Spanish (see laplace.ngo). The EALTA group's goal is for migrant and refugee language programs to be able to use these tests for placement purposes. These projects sound promising, and seem to address Shohamy's (in Lazaraton 2010) warning against being wooed too much by the 'magic' of cloze testing: these projects proceed with caution and care in using more robust assessments with multiple individual tests in a computer-adaptive framework. Like C-tests, cloze testing is good because of its practicality, but the score interpretations from any one cloze text should not be overreaching. Creating a series of difficulty-overlapping cloze test passages for a computer-adaptive series of cloze tests will provide more, articulated information, and will pinpoint on a broader proficiency scale language learners' ability to read in the target language, which provides probable insight on the test-taker's vocabulary level and their range of grammatical knowledge.

Acknowledgements

This research was brainstormed initially at an eye-tracking and language assessment pre-conference workshop in East Lansing, Michigan, held before the 2014 meeting of the Midwest Association of Language Testers (MwALT, mwalt.msu.edu). The authors, who were the workshop leader and two of the attendees, respectively, would like to thank the following workshop attendees for their insights and help with earlier versions of this manuscript: Suthathip (Ploy) Thirakunkovit, Katie Weyant and Daniel Walter. The authors would also like to thank Michigan Language Assessment for its contributions to the overall project.

References

Alderson, C (1979) The cloze procedure and proficiency in English as a foreign language, *TESOL Quarterly* 13 (2), 219–227. doi:10.2307/3586211

American Council on the Teaching of Foreign Languages (ACTFL) (2012) *The ACTFL Proficiency Guidelines,* Alexandria: American Council on the Teaching of Foreign Languages, available online: www.actfl.org/educator-resources/actfl-proficiency-guidelines

American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME] and Joint Committee on Standards for Educational and Psychological Testing (2014) *Standards for Educational and Psychological Testing*, Washington, D.C.: American Educational Research Association.

Bachman, L F (1985) Performance on cloze tests with fixed-ratio and rational deletions, *TESOL Quarterly* 19 (3), 535–556. doi:10.2307/3586277

Brown, J D (2013) My twenty-five years of cloze testing research: So what?, *International Journal of Language Studies* 7 (1), 1–32, available online: http://www.ijls.net/sample/71-1.pdf

Brown, J D, Trace, J, Janssen, G and Kozhevnikova, L (2016) How well do cloze items work and why?, in Gitsaki, C and Coombe, C (Eds) *Current Issues in Language Evaluation, Assessment, and Testing*, Newcastle: Cambridge Scholars Publishing, 2–39.

Brunfaut, T and McCray, G (2015) *Looking into test-takers' cognitive processes whilst completing reading tasks: A mixed-method eye-tracking and stimulated recall study*, ARAGs Research Reports Online, Volume AR/2015/001, London: British Council, available online: www.britishcouncil.org/sites/default/files/brunfaut-and-mccray-report_final.pdf

Chavez-Oller, M A, Chihara, T, Weaver, K A and Oller, J W (1985) When are cloze items sensitive to constraints across sentences?, *Language Learning* 35 (2), 181–206. doi:10.1111/j.1467-1770.1985.tb01024.x

Ferlazzo, L (2012) *The Best Tools for Creating Clozes (Gap-Fills)*, available online: larryferlazzo.edublogs.org/2012/04/30/the-best-tools-for-creating-clozes-gap-fills/

Fotos, S S (1991) The cloze test as an integrative measure of EFL proficiency: A substitute for essays on college entrance examinations?, *Language Learning* 41 (3), 313–336. doi:10.1111/j.1467-1770.1991.tb00609.x

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

Kincaid, J P, Fishburne Jnr, R P, Rogers, R L and Chissom, B S (1975) *Derivation of New Readability Formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy Enlisted Personnel*, Research Branch Report 8–75, Millington: Chief of Naval Technical Training, available online: stars.library.ucf.edu/cgi/viewcontent.cgi?article=1055&context=istlibrary

Kobayashi, M (2002) Cloze tests revisited: Exploring item characteristics with special attention to scoring methods, *The Modern Language Journal* 86 (4), 571–586. doi:10.1111/1540-4781.00162

Kremmel, B and Schmitt, N (2016) Interpreting vocabulary test scores: What do various item formats tell us about learners' ability to employ words?, *Language Assessment Quarterly* 13 (4), 377–392. doi:10.1080/15434303.2016.1237516

Lazaraton, A (2010) From cloze to consequences and beyond: An interview with Elana Shohamy, *Language Assessment Quarterly* 7 (3), 255–279. doi:10.1080/15434301003792815

Markham, P L (1985) The rational deletion cloze and global comprehension in German, *Language Learning* 35 (3), 423–430. doi:10.1111/j.1467-1770.1985.tb01085.x

McCray, G and Brunfaut, T (2018) Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking, *Language Testing* 35 (1), 51–73. doi:10.1177/0265532216677105

McKenna, M C and Layton, K (1990) Concurrent validity of cloze as a measure of intersentential comprehension, *Journal of Educational Psychology* 82 (2), 372–377. doi:10.1037/0022-0663.82.2.372

Messick, S (1990) Validity of test interpretation and use, *ETS Research Report Series 1990* 1, 1,487–1,495. doi:10.1002/j.2333-8504.1990.tb01343.x

Messick, S (1995) Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning, *American Psychologist* 50 (9), 741–749. doi:10.1002/j.2333-8504.1994.tb01618.x

Messick, S (1996) *Validity and Washback in Language Testing*, Educational Testing Service Research Report, Princeton: Educational Testing Service.

Oller, J W (1973) Cloze tests of second language proficiency and what they measure, *Language Learning* 23 (1), 105–118. doi:10.1111/j.1467-1770.1973.tb00100.x

Polio, C (2012) Replication in published applied linguistics research: A historical perspective, in Porte, G (Ed) *Replication Research in Applied Linguistics*, Cambridge: Cambridge University Press, 47–91.

Reichle, E D, Pollatsek, A, Fisher, D L and Rayner, K (1998) Toward a model of eye movement control in reading, *Psychological Review* 105 (1), 125–157. doi:10.1037/0033-295X.105.1.125

Sadeghi, K (2014) Phrase cloze: A better measure of reading?, *The Reading Matrix* 14 (1), 76–94, available online: www.readingmatrix.com/articles/april_2014/sadeghi.pdf

Scholastic (2012) *Quick Cloze Passages for Booking Comprehension*, Grades 4–6, New York: Scholastic Corporation.

Shohamy, E (1981) The cloze procedure and its applicability for testing Hebrew as a foreign language, *Hebrew Annual Review* 5, 101–114, available online: kb.osu.edu/handle/1811/58634

Taylor, W L (1953) "Cloze procedure": A new tool for measuring readability, *Journalism & Mass Communication Quarterly* 30 (4), 415–433. doi:10.1177/107769905303000401

Tobii Studio (2016) *Tobii Studio User's Manual*, Version 3.4.5, available online: www.staff.universiteitleiden.nl/binaries/content/assets/sociale-wetenschappen/faculteitsbureau/solo/research-support-website/software/tobii-pro-studio-user-manual_3.4.5_08082019.pdf

Trace, J, Brown, J D, Janssen, G and Kozhevnikova, L (2017). Determining cloze items difficulty from item and passage characteristics across learner backgrounds, *Language Testing* 34 (2), 151–174. doi:10.1177/0265532215623581

Tremblay, A (2011) Proficiency assessment standards in second language acquisition research: "Clozing" the gap, *Studies in Second Language Acquisition* 33 (3), 339–372. doi:10.1017/S0272263111000015

Tremblay, A and Garrison, M D (2010) Cloze tests: A tool for proficiency assessment in research on L2 French, in Prior, M T, Watanabe, Y and Lee, S-K (Eds) *Selected Proceedings of the 2008 Second Language Research Forum*, Somerville: Cascadilla Press, 73–88.

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

Willingham, W W and Cole, N (1997) *Gender and Fair Assessment*, Mahwah: Lawrence Erlbaum Associates.

Winke, P (2024) *Supplemental materials: What does the cloze test really test? A cognitive validation of a French cloze test with eye tracking and interview data*, Open Science Framework. doi:10.17605/OSF.IO/2WRPK

Yamashita, J (2003) Processes of taking a gap-filling test: Comparison of skilled and less skilled EFL readers, *Language Testing* 20 (3), 267–293. doi:10.1191/0265532203lt257oa

Zou, D (2017) Vocabulary acquisition through cloze exercises, sentence-writing and composition-writing: Extending the evaluation component of the involvement load hypothesis, *Language Teaching Research* 21 (1), 54–75. doi:10.1177/1362168816652418

Appendix 1: Summary of test scores and survey responses

ID	French level	Gender	Age	Year in college	Studied/ lived abroad	Self-assessed: Reading ability	Self-assessed: Comprehensibility of test passage	Self-assessed: Difficulty of the passage
1	Graduate NS	M	25+	MA	Lived: 1 yr. +	5	10	2
2	Instructor NS	F	25+	Post PhD	Lived: 1 yr. +	5	10	3
3	Instructor	M	25+	Post MA	Lived: 1 yr. +	5	10	2
4	300	M	19	Senior	No	3	5	3
5	400	M	22	Senior	Studied: 6 weeks	3	7	3
6	400	F	20	Junior	No	3	6	4
7	300	M	20	Sophomore	No	2	5	3
8	200	M	19	Sophomore	No	3	5	3
9	200	F	20	Sophomore	No	3	7	3
10	300	F	21	Senior	No	3	4	4
11	300	F	20	Sophomore	No	3	6	2
12	200	F	19	Freshman	No	4	8	3
13	400	M	23	Senior	No	2	3	2
14	400	F	20	Junior	No	3	7	3
15	300	F	20	Senior	Studied: 6 weeks	3	8	3
16	200	M	25+	Senior	No	3	5	3
17	100	M	21	Senior	No	3	6	4
18	100	F	25+	MA	No	1	7	4
19	100	F	18	Sophomore	No	2	5	2
20	200	F	22	Senior	Studied: 6 weeks	3	6	4
21	100	M	19	Freshman	No	2	3	4
22	100	F	19	Freshman	No	1	1	3

Main idea of test passage	Language skills measured by the test	Appropriacy of the test	Prior test taking	Total score (max. = 45)	Test time (min:sec)	Test time (seconds)
Yes	Reading	Yes	No	45	15:21	921
Yes	Grammar Guessing	Yes	No	42	7:43	463
Yes	Vocab	Yes	No	33	15:07	907
Yes	Grammar Vocab	No	No	25	13:03	783
Yes	Grammar	Yes	No	25	62:01	3,662
Yes	Grammar	Yes	No	21	23:56	1,436
Yes	Grammar	Yes	No	19	12:53	773
Yes	Grammar Vocab	Unsure	Yes	18	26:18	1,578
Yes	Reading	Yes	No	18	15:00	900
Yes	Grammar	Yes	No	18	21:49	1,309
Yes	Comprehension	Yes	No	17	45:29	2,729
Yes	Grammar	Unsure	No	16	15:53	953
Yes	Grammar	Yes	No	15	44:24	2,664
Yes	Science Knowledge	Yes	No	14	23:31	1,411
Yes	Reading Grammar	Yes	No	12	11:45	705
Yes	Reading Vocab	Unsure	No	9	44:11	2,651
Yes	Grammar Vocab	No	No	8	9:52	592
Yes	Memory Grammar	No	No	8	15:15	915
Yes	Grammar Vocab	No	No	3	8:25	505
Yes	Grammar Vocab	No	No	3	10:17	617
Yes	Grammar Vocab	No	Yes	2	13:37	817
No	Vocab	No	No	0	3:41	221
Average (SD)				16.45 (11.46)	20:45 (870.861)	1,251

Appendix 2: Additional post-session interview questions (for the one-on-one semi-structured interviews for all participants)

1. Have you ever taken this type of test before?
2. Did you get the main idea of the passage?
3. In your opinion, what is the topic or main idea of the passage?
4. What process did you use to determine the correct answers? (That is, what did you do to figure out correct answers?)
5. What (if any) test-taking strategies did you use to fill in the blanks?
6. Did you think the test was appropriate for you?
7. In your opinion, what language skill(s) do you think was measured by the test?
8. Do you have any further comments on the passage, its contents, or the procedure?

3 The use of eye tracking in validating L2 listening assessments

Ruslan Suvorov
University of Western Ontario, Canada

Abstract

Despite the importance of process-oriented approaches to validation that entail gathering validity evidence based on the analysis of individual response processes (Wu and Stone 2015), their use in validation research has been fairly limited and done primarily via concurrent and retrospective verbal reports. Given the limitations of verbal reports that are prone to veridicality and reactivity risks (Bowles 2010), some researchers have employed eye-tracking technology to investigate test-takers' response processes in L2 assessment contexts as part of validation research (Bax and Chan 2016). With listening still being the most under-researched L2 skill, validation studies that gather validity evidence based on response processes are particularly scarce in the context of L2 listening assessment (Aryadoust 2019b) and sorely needed for understanding L2 listeners' cognitive processes and test-taking strategies underlying their performance.

To address this gap in research, this chapter introduces two studies that illustrate the potential of eye tracking for gathering response process data as a source of cognitive validity evidence in the context of L2 listening assessment. Study 1 addresses the question of whether visuals should be integrated into L2 listening tests by exploring test-takers' viewing behaviour during their interaction with two video types in a video-mediated listening test. Study 2 examines test-taking strategies used by L2 listeners when completing three types of listening item, as well as the effect of test-wiseness strategies on their test scores. The results of both studies are discussed with regard to their implications for process-based validation research on assessing L2 listening.

Introduction

The importance of process-oriented approaches to validation, which entail gathering validity evidence based on the analysis of individual response processes, is widely acknowledged (e.g., American Educational Research Association [AERA], American Psychological Association [APA], National

Council on Measurement in Education [NCME] and Joint Committee on Standards for Educational and Psychological Testing 2014, Messick 1995, Wu and Stone 2015). Unlike product-oriented approaches that focus on assessment outcomes (such as test-takers' oral or written responses) and rely predominantly on the use of statistical methods to analyse test scores and validate their interpretations and uses, process-oriented approaches to validation can evince test-takers' cognitive processes and test-taking strategies and provide valuable insights into the constructs measured by such tests (Anderson, Bachman, Perkins and Cohen 1991, Schmitt, Ng and Garras 2011).

Despite the growing recognition of the need to expand research on the processes and strategies underlying second language (L2) learners' responses to test items, validation research utilising process-oriented approaches has been fairly limited (Cohen 2014, Ercikan and Pellegrino (Eds) 2017, Padilla and Benítez 2014, Wu and Stone 2015, Zumbo and Hubley (Eds) 2017) and done primarily via concurrent and retrospective verbal reports such as interviews (Phakiti 2003), questionnaires (Kashkouli and Barati 2013) and think-aloud protocols (Cohen and Upton 2007, Plakans 2009). Given the limitations of concurrent and retrospective verbal reports that are prone to over-reporting, under-reporting, or distortion of self-reported thoughts, behaviours, and cognitive processes (Bowles 2010), some researchers have begun to employ eye-tracking technology, which is capable of measuring the viewing behaviour and visual attention of L2 learners objectively, to investigate test-takers' response processes in the L2 assessment context as part of validation research (e.g., Bax 2013, Bax and Chan 2016, Brunfaut and McCray 2015). With listening still being recognised as the most under-researched L2 skill (Harding 2012), validation studies that aim at gathering validity evidence based on test-takers' response processes are particularly scarce in the context of L2 listening assessment (e.g., Aryadoust 2019b, Winke and Lim 2014) and sorely needed for understanding L2 listeners' cognitive processes and the test-taking strategies underlying their performances and test scores, as well as the constructs measured by such tests.

To address this gap in research, this chapter introduces two eye-tracking studies that illustrate the potential of eye tracking for gathering response process data as a source of cognitive validity evidence in the context of L2 listening assessment. Specifically, Study 1 attempts to tackle the unresolved issue of whether visuals should be integrated into L2 listening tests by utilising eye tracking to explore the viewing behaviour of test-takers during their interaction with two video types in a video-mediated L2 academic listening test. Meanwhile, eye-tracking Study 2 aims at exploring the types of test-taking strategies used by L2 learners during their completion of three types of listening items from the Michigan English Test (MET), as well as the effect of test-wiseness strategies on their listening test scores. To foreground

the two studies, the chapter begins with a concise overview of research on L2 listening, discusses the value and importance of cognitive validity in Weir's (2005) socio-cognitive framework, and introduces the growing body of studies that use eye tracking for process-oriented validation research, including the validation of L2 listening assessments.

Literature review

Brief overview of L2 listening

As one of the four language skills, listening is widely regarded as a complex, multifaceted skill that plays an essential role in L2 acquisition (Field 2008, Rost 2002, Vandergrift 2012). Due to its complex and transient nature, listening still remains a relatively under-investigated (Harding 2012), less well understood (Morley 2001, Vandergrift 2010), and somewhat less commonly taught L2 skill, partly due to the fallacious expectation by some educators that it would develop automatically as part of the speaking practice (Seo, Taherbhai and Frantz 2016). It is therefore not surprising that the assessment of listening is challenging and complex to perform (Buck 2001, Field 2013) due to the interplay of both *internal* factors that include listeners' physiological characteristics, cognitive processes, and affective states, and *external* contextual factors such as the setting, the purpose of the listening task, and speaker-related characteristics such as the degree of accentedness (Goh and Aryadoust 2016, Taylor and Geranpayeh 2011, Vandergrift 2010).

The complexity of listening is highlighted in ongoing attempts to explain the processes involved in listening comprehension and define and conceptualise this multidimensional skill that is comprised of various listening sub-skills (e.g., low-level listening skills such as word recognition, and high-level listening skills such as inferencing; see, for instance, Buck 2001, Field 2008, Taylor and Geranpayeh 2011). Given that the goal of developing a single, universally accepted definition of listening is most likely unattainable due to its complex nature, some researchers (e.g., Bodie, Janusik and Välikoski 2008) have proposed using multiple definitions depending on the type of listening and the underlying cognitive processes. Existing studies of L2 listening have used one of the following three models to investigate and explain the process of listening comprehension: the bottom-up model, the top-down model, and the interactive model (Flowerdew and Miller 2010). According to the bottom-up model, a listener constructs meaning from an acoustic signal by first processing the smaller units of the input such as phonemes and words and then moving to decode discourse-level units such as sentences and paragraphs. Unlike the bottom-up model, the top-down model suggests that processing of the incoming aural input starts with an activation of listeners' background knowledge (schemata) that allows them

to anticipate what they will hear before decoding phoneme-level units. Finally, the interactive model of listening comprehension posits that the activation of prior knowledge and the processing of phonological, syntactic, and pragmatic information occur simultaneously and in an interactive manner.

Regardless of which model is used to explain the process of listening, there appears to be a consensus that listening in general is an internalised, active skill (Rost and Wilson 2013) employed for both a unidirectional processing of the input (e.g., while listening to an academic lecture) and a reciprocal (two-way) processing of the input that normally occurs during an interaction between two interlocutors (Lynch 2011). Listening entails multitudinous processes in cognitive, affective, and behavioural domains, including lower-level processes such as input decoding and parsing and higher-level processes such as incremental meaning building and discourse construction (Field 2008, 2013, Rost 2014, Seo et al 2016). The fact that these processes cannot be observed directly, however, poses a number of challenges for their assessment and validation practices.

Traditionally, the assessment of L2 listening comprehension has focused on the *product* of listening – such as the scores derived from examinees' responses to test items, oftentimes presented in a multiple-choice format that has dominated large-scale testing (Lee and Winke 2012) – rather than the *process* of listening (Field 2013). As a result, the validation of listening assessments has heavily relied on the use of product-oriented, or outcome-based, approaches that predominantly utilise statistical methods to analyse test scores based on *a posteriori* assumptions (e.g., Ginther 2002, Londe 2009, Wagner 2010b), and validate their interpretations and uses as part of the process of gathering the evidence of cognitive validity (Taylor 2013).

Cognitive validity and the socio-cognitive framework

Cognitive validity (previously known as theory-based validity; see, for instance, Weir 2005) refers to the extent to which individual assessment tasks elicit the evidence of L2 learners' cognitive processes and mental abilities that resemble those occurring in the relevant real-world listening situations (Field 2013, Taylor 2013). Alongside other types of validity, cognitive validity plays a prominent role in Weir's (2005) socio-cognitive framework, which is designed to validate language assessments. What sets this framework apart from other important frameworks used for validation purposes such as the Assessment Use Argument (Bachman 2005, Bachman and Palmer 2010) and the argument-based approach (Chapelle 2011, Chapelle, Enright and Jamieson (Eds) 2008, Kane 1992, 2006, 2010) is the explicit emphasis on the cognitive processes underpinning test-takers' performances, as well as their interaction with contextual factors and the process of scoring.

In light of the interconnectedness between cognitive validity and scoring aspects of validity advocated by the socio-cognitive framework, Weir (2005) warns against the over-reliance on psychometric properties of the product of assessment (i.e., test scores) to make *a posteriori* inferences about the processes underlying each individual test-taker's performance. Indeed, outcome-oriented assessment and validation do not yield any direct evidence of the *processes* underlying test-takers' performances, and similar concerns about excessive trust and reliance on statistical modelling in validation work have been expressed by other researchers too (e.g., Launeanu and Hubley 2017b, Mislevy 2009). For instance, when discussing the validity of inferences made on the basis of conventional test score interpretations informed by Item Response Theory (IRT), Mislevy (2009) reminded that 'IRT characterizations of students and items [...] are clearly simplifications, and they say nothing about the processes by which students answer items' (2009:3).

By acknowledging the value and importance of cognitive validity in validation research, Weir's (2005) socio-cognitive framework signals the need to garner direct empirical evidence of physiological, psychological, and experiential characteristics of test-takers, as well as evidence about the test-takers' actual processing of the test tasks (e.g., their use of metacognitive strategies or strategic competence). In the context of L2 listening assessment, such evidence would contain information about response processes that L2 test-takers employ to complete listening assessment tasks.

Response processes as a source of validity evidence

Response processes can be defined as 'thought processes, strategies, approaches, and behaviours of examinees when they read, interpret, and formulate solutions to assessment tasks' (Ercikan and Pellegrino (Eds) 2017:2). In recent years, there has been a burgeoning interest in validation research that produces validity evidence based on response processes (see, for instance, the edited volumes by Ercikan and Pellegrino (Eds) 2017, Zumbo and Hubley (Eds) 2017). The importance of response-process validity evidence is also well documented in authoritative publications such as the *Standards for Educational and Psychological Testing* (AERA et al 2014), where response processes are cited as one of the five sources of validity evidence along with test content, internal structure, relations to other variables, and consequences of testing.

Collecting evidence based on response processes has a number of benefits for evaluating the validity of test scores and their interpretations. In particular, response-process data can yield insights into and increase an understanding of a) the nature of test-takers' interaction with assessment tasks and test-takers' response patterns, including the extent to which these

occur in the expected ways (Ercikan and Pellegrino (Eds) 2017, Kane and Mislevy 2017); b) the types of strategies (including test management and test-wiseness, or test-deviousness, strategies; Cohen 2014) that test-takers utilise when interacting with and answering each specific item; c) the temporal characteristics of test-takers' behaviour such as the steps they take to complete a specific task, their order, and the amount of time they spend on each task and step (van der Linden 2009); and d) the fit between the construct and test-takers' response processes, including the extent to which 'capabilities irrelevant or ancillary to the construct may be differentially influencing test-takers' test performance' (AERA et al 2014:15).

The importance of understanding L2 learners' response processes has been also acknowledged by both researchers and practitioners working in the area of L2 listening (e.g., Field 2008, Vandergrift 2010). According to Vandergrift (2010), a process-oriented approach to investigating L2 listening can 'provide opportunities for participants to reveal, or researchers to uncover, listener decision-making processes during comprehension' (2010:165) – information that cannot be obtained or inferred solely from the analysis of test scores and that is critical for gaining a better understanding of the cognitive processes underlying L2 learners' listening comprehension and making more informed interpretations of their test scores. Understanding these cognitive processes in L2 listening assessment contexts requires an examination of the complex interplay between test-takers' listening processes (e.g., extracting meaning from auditory and/or visual stimuli) and reading processes (e.g., reading and responding to the listening test items presented in a written format) that are believed to share many commonalities as postulated by Aryadoust's (2019a) integrated cognitive theory of reading and listening.

Despite the recognition of the need to investigate the processes and strategies underlying L2 learners' responses to test items, validation research utilising process-oriented approaches has been fairly limited (Cohen 2014, Ercikan and Pellegrino (Eds) 2017, Lee and Winke 2017, Padilla and Benítez 2014, Wu and Stone 2015, Zumbo and Hubley (Eds) 2017). For instance, existing studies of test-taking strategies have relied primarily on the use of concurrent and retrospective verbal reports such as interviews (Phakiti 2003), questionnaires (Kashkouli and Barati 2013), and think-aloud protocols (Cohen and Upton 2007, Plakans 2009), with the bulk of research in this area having been done in the context of L2 reading assessment. However, the use of methodologies that rely on the collection and analysis of self-reported data is somewhat problematic because they are prone to reactivity and non-veridicality threats (Bowles 2010). In other words, when asked to verbalise their cognitive processes – either introspectively or retrospectively – respondents tend to forget the details of their thoughts during the initial task (which may lead to under-reporting or even fabrication of information) and get affected by their own verbalisations in a way that changes their thought processes.

While still providing valuable data, self-reported methods have a limited potential to produce consistent and cogent evidence of L2 learners' response patterns and engagement with assessment tasks, with the general sentiment being that 'verbal reports are unlikely to faithfully reflect everything that learners are attending to and aware of' (Robinson, Mackey, Gass and Schmidt 2012:261). Taking into consideration that self-reported methods can yield mainly the data about *inferred* response processes (Launeanu and Hubley 2017a), there is a growing demand for the use of more sensitive and objective measures – such as eye tracking (Hubley and Zumbo 2017, Oranje, Gorin, Jia and Kerr 2017, Winke and Lim 2014) – that are capable of affording cogent evidence of *observed* response processes. It is believed that such measures can shed light on some of the previously unknown, or lesser known, processes underlying test-takers' performances during L2 assessments, especially the performances requiring the use of receptive skills such as listening and reading that are more difficult to directly observe and measure.

Eye tracking in L2 assessment

Eye tracking refers to the use of specialised technology for recording the viewer's eye movements during the completion of a specific task and subsequently calculating and analysing various metrics associated with the eye-movement data (Duchowski 2007, Holmqvist et al 2011). Recent advances in hardware and software, as well as the falling costs, have led to eye tracking becoming more accessible to researchers interested in leveraging this non-intrusive technology to investigate different aspects of viewers' behaviour and cognition. Although eye movements can only provide direct evidence of what the viewers look at and not what they think about, eye tracking has been widely used for investigating cognitive processes based on the eye-mind hypothesis (Just and Carpenter 1980). According to this hypothesis, there is a strong relationship between gaze (i.e., what one looks at) and mental processing (what one thinks about), which implies that, for instance, a prolonged gaze at a specific area of the visual field may serve as an indication that the viewer is cognitively involved with that area. In light of the eye-mind hypothesis, eye tracking is justifiably deemed to be a valuable method for gathering evidence of not only behavioural aspects of performance (e.g., the order in which the options on a multiple-choice item are viewed), but also the underlying cognitive processes.

Eye tracking has been around for many decades and has firmly established itself in psychology and psycholinguistic research (Rayner 2009), and more recently also in second language acquisition (SLA) research, albeit to a smaller extent (see, for instance, Conklin and Pellicer-Sánchez 2016, Conklin, Pellicer-Sánchez and Carrol 2018, Godfroid and Winke 2015, Godfroid, Winke and Gass (Eds) 2013, Winke, Gass and Sydorenko 2013).

Although the application of this methodology in the field of second language testing and assessment has been a relatively recent development, it is indisputably gaining strong momentum among language testing specialists and researchers (e.g., Bax 2013, Bax and Chan 2016, Bax and Weir 2012, Brunfaut and McCray 2015, McCray and Brunfaut 2018, Winke et al 2018, Winke and Lim 2014, 2015). Most eye-tracking research on L2 assessment pertains to reading, although some studies have also addressed other skills such as speaking (e.g., Lee and Winke 2017), writing (e.g., Révész, Michel and Lee 2017, Winke and Lim 2015), and listening (e.g., Aryadoust 2019b, Batty 2017, Holzknecht et al 2017, Winke and Lim 2014).

In the studies of L2 reading that aimed at gathering cognitive validity evidence, eye tracking has been utilised to investigate test-takers' cognitive processes during their completion of test items from the International English Language Testing System (IELTS) Reading test (Bax 2013), *Cambridge English: Advanced* (now C1 Advanced) Reading test (Bax and Weir 2012), General English Proficiency Test (GEPT) Reading test (Bax and Chan 2016), the reading component of the Aptis test (Brunfaut 2016, Brunfaut and McCray 2015), and PTE (Pearson Test of English) Academic (McCray and Brunfaut 2018). These investigations were based both on the qualitative analysis of eye–movement recordings and on the quantitative analysis of eye-tracking metrics such as total fixation duration, the number of forward saccades and regressions, and the total number of fixations and visits to a specific area of interest (AOI). One of the major implications of this line of research is that, compared to verbal protocol methods, eye-tracking methods are capable of providing more insights into cognitive processing during L2 reading, with some of these studies (e.g., Bax and Chan 2016, Brunfaut and McCray 2015) advocating the importance of combining eye tracking with verbal reports such as stimulated recalls for the purposes of data triangulation.

In research on L2 listening assessment, the application of eye-tracking methodology for gathering response process validity evidence has been scarce. Winke and Lim (2014), for instance, investigated the extent to which test-wiseness and test-taking anxiety affected L2 learners' performance on the IELTS Listening test by analysing the data from eye-movement recordings (i.e., total fixation duration and the number of fixations per AOI) and stimulated recall interviews. The eye-movement data afforded some useful evidence of test-takers' response processes, namely, that high scorers demonstrated faster gaze fixations on key words compared to low scorers, and that highly anxious test-takers tended to spend more time reading the test instructions and processing key words. Despite the benefits that the application of eye tracking brought to this study, the authors acknowledged that they 'were limited in [their] ability to analyze the data empirically because [they] had to, in a sense, invent the wheel' (Winke and Lim 2014:20).

This was primarily due to the scarcity of eye-tracking studies in the field of language testing and the lack of established methods for analysing the eye-movement data that the authors could avail themselves of.

In a more recent eye-tracking study, Aryadoust (2019b) examined a) how the reading of multiple-choice and matching items by 28 listeners changes before and during two hearings of listening prompts in an online while-listening performance test, and b) how these listeners engage in answer changing during the two hearings. The results of cross-correlation analysis and multivariate analysis of variance of the eye-tracking data evinced that the listeners' reading patterns changed significantly across the four rounds of item reading, with the item reading during while-listening being significantly more cognitively demanding than the item reading during pre-listening. Additionally, the investigation of answer-changing patterns detected both correct-to-incorrect and incorrect-to-correct answer-changing patterns, with the latter pattern being more prevalent. In his discussion of the findings, Aryadoust (2019b) stated that '[t]his dynamicity invites an inevitable question about the validity of test scores that tend to mask these mechanisms in the assessment of listening' (2019b:20), suggesting that the harnessing of eye-tracking technology for scrutinising test-takers' response processes is critical for the validation of L2 listening assessments as it provides validity evidence that traditional test scores are unable to provide.

To contribute to this inchoate but critical body of research, the rest of this chapter describes two studies aimed at leveraging eye tracking to collect response process evidence in support of the cognitive validity of L2 listening tests, with the first study being the summary of the studies reported in Suvorov (2015, 2018b) and the second study being part of a larger project reported in Suvorov (2018a).

Study 1: Exploring the use of visual information during video-mediated L2 listening assessment

Background

Visuals have traditionally been viewed as a critical component of L2 listening comprehension, yet their use in L2 listening tests and role in the construct of L2 listening have been the subject of contentious debate (Batty 2015, Buck 2001, Ockey 2007, Wagner 2007). A large body of work in this area comprises comparative studies that attempted to draw conclusions about the effect of the visual medium (mostly video, but also pictures) on L2 test-takers' performances based on the analysis of their scores from audio-only vs. multimedia-enhanced listening assessment tasks. The findings have been very mixed, indicating that the inclusion of visuals in L2 listening assessment tasks could have a facilitative effect (e.g., Ginther 2002, Shin 1998, Sueyoshi and

Hardison 2005, Wagner 2010b, 2013), a debilitating effect (e.g., Pusey and Lenz 2014, Suvorov 2009), or a neutral effect (e.g., Batty 2015, Coniam 2001, Cubilo and Winke 2013, Gruba 1993, Londe 2009) on test-takers' scores.

These differing results can be attributed to a plethora of variables such as different types of visuals, text types, item types, test-takers' proficiency levels, and administration procedures used in the tests, thereby rendering the studies disparate. With a few exceptions (e.g., Ginther 2002), comparative studies have traditionally focused on exploring the audio-visual dichotomy without factoring in different visual types such as context visuals (i.e., visuals that provide information about the setting and the speaker) and content visuals (i.e., visuals that provide information related to the content of the auditory stimulus) (Bejar, Douglas, Jamieson, Nissan and Turner 2000). In light of the differences between context and content visuals with regard to their semantic load, investigations of how L2 test-takers interact with these visual types appear to be of paramount importance for understanding the potentially differing effects of these visuals on the test-takers' listening comprehension and test performances.

Another problematic issue pervading all comparative studies appears to be a notable disregard for test-takers' viewing behaviour and interaction with the visual input during the assessments. Without gathering the direct evidence of test-takers' viewing behaviour, some of whom might not even look at the visuals while taking a multimedia-enhanced listening test (Ockey 2007, Wagner 2007, 2010a), it is easy to question the validity of claims about the effect of visuals on L2 test-takers' listening comprehension that are made solely on the basis of test scores.

To address this gap in the literature, Study 1 aimed at providing observed evidence of L2 learners' viewing behaviour during a video-mediated L2 listening test by leveraging eye tracking to record their interactions with two video types: context videos and content videos. This study intended to answer the following research questions (RQs):

1. To what extent do L2 test-takers watch context videos differently from content videos?
2. What aspects of visual information in the two video types do L2 test-takers find helpful and/or distracting for their listening comprehension and test performance?

Study 1 overview

In Study 1, EyeTech Vision Tracker 2 (80 hertz (Hz) data sampling rate, 0.5 degree of visual angle accuracy, 65–100cm operating range, 1,680 × 1,050 display) was used to record the eye movements of 33 participants while they were completing a video-based academic listening test. The participants

comprised non-native English-speaking test-takers: 25 undergraduate students from English as a second language (ESL) listening courses, and eight graduate students in an applied linguistics programme at a large university in the Mid-west of the USA. The video-based academic listening test that was developed for the larger project (Suvorov 2013) consisted of six short (3–4 minutes long) video clips and 30 test items presented in a four-option multiple-choice format, with five test items following each video clip. The video clips were taken from authentic university lectures, with three clips classified as context videos and the other three classified as content videos. The test was delivered online and lasted for approximately 45 minutes. Paper-based note-taking was allowed during the test. After completing the test, each test-taker was invited to take part in cued retrospective reporting (Van Gog, Paas, Van Merriënboer and Witte 2005), a data collection method that entails showing the participants the recordings of their eye movements that were made during their completion of the initial task and asking them to verbalise their response processes (namely, their viewing patterns and use of visual information from the videos) while being cued by these eye-movement recordings. All verbalisations were audio-recorded for a subsequent analysis.

To determine whether test-takers exhibited different viewing patterns depending on the type of video (RQ1), the eye-tracking data were analysed by calculating three metrics – fixation rate, dwell rate, and total dwell time – and comparing them for the two video types with the help of paired-samples *t*-tests. Fixation rate (i.e., the number of eye fixations per second) was used as a measure of the semantic importance of each video clip. Dwell rate (i.e., the number of visits into a specific AOI per minute) was used as a measure of the importance of the video AOI (i.e., the area of the screen in which each video was played) for the completion of test items. The total dwell time (i.e., the percentage of time that a test-taker spent looking at the video AOI) was used as a measure of test-takers' interest in the AOI-confined visual or its informativeness. The audio-recorded verbal data were transcribed and analysed qualitatively to identify the aspects of visual information that the test-takers found helpful or distracting for their listening comprehension and for answering each individual test item (RQ2).

In response to RQ1, the results of the quantitative analysis revealed a statistically significantly higher fixation rate for content videos (M = .87, SD = .42) than for context videos (M = .71, SD = .40), $t(32) = 4.73$, $p < .01$, eta squared = .41 (moderate effect size). The percentage of the total dwell time was also statistically significantly higher for content videos (M = 57.99, SD = 19.79) than for context videos (M = 50.70, SD = 22.49), $t(32) = 5.02$, $p < .01$, eta squared = .44 (moderate effect size). With regard to the dwell rate, however, there was no statistically significant difference between content videos (M = 29.40, SD = 15.49) and context videos (M = 29.07, SD = 17.26), $t(32) = .38$, $p = .71$, eta squared = .01 (small effect size).

In response to RQ2, the results of the retrospective verbal report data analysis demonstrated that the test-takers attended to different visual aspects of the two video types during the listening test. In particular, the findings revealed two main groups of visual information – namely, speaker-related and lecture-related aspects – that the L2 test-takers perceived as having a facilitative or a distracting effect on their listening comprehension and test performance. Speaker-related aspects referred to essential attributes of the speaker in each video (e.g., gestures, facial expressions, and body movements), whereas lecture-related aspects comprised all other, non-speaker-related visual elements of each video (e.g., visual aids and textual information) that were relevant to the content of the auditory stimulus. Overall, the participants reported focusing on speaker-related aspects more than on lecture-related aspects (160 references vs. 124 references, respectively), but found lecture-related aspects to be overwhelmingly more helpful for their listening comprehension and test performance than speaker-related aspects (123 references vs. 25 references, respectively).

Study 2: Examining test-taking strategies during the completion of L2 listening test items

Background

Test-taking strategies are deemed to be instrumental for construct validation and regarded as a valuable source of validity evidence based on response processes (Cohen 2007, Schmitt et al 2011, Wu and Stone 2015, Wu and Zumbo 2017). In accordance with Cohen's (2014) classification, test-taking strategies can be divided into *test management* strategies that entail response processes relevant to the construct measured by the test, and *test-wiseness* (or test-deviousness) strategies (e.g., guessing) that introduce construct-irrelevant variance.

While investigations of test-taking strategies are abundant (e.g., Cohen and Upton 2007, Phakiti 2003, Plakans 2009, Schmitt et al 2011, Yamashita 2003), their application for language test validation purposes has been scant. Data elicitation in the vast majority of existing studies on L2 test-taking strategies has been carried out via verbal report methodologies such as think-aloud protocols (Anderson et al 1991, Cohen and Upton 2007, Plakans 2009), retrospective protocols (Sasaki 2000), questionnaires (Kashkouli and Barati 2013), introspective interviews (Schmitt et al 2011), and retrospective interviews (Phakiti 2003, Plakans 2009). Although verbal reports can afford certain insights into L2 learners' use of test-taking strategies (Wu and Zumbo 2017), their potential for eliciting detailed information from test-takers is limited. Complementing verbal report methods with more direct measures of test-takers' response processes, such as the recordings of their eye

movements, can arguably result in more detailed and compelling evidence of the nature and extent of the test-taking strategies employed during the assessment (Brunfaut and McCray 2015).

In order to address the lack of research that highlights the combined potential of eye tracking and verbal reports for collecting information about test-taking strategies in L2 listening assessments, Study 2 aimed at exploring the types of test-taking strategies used by L2 learners and investigating the effect of test-wiseness strategies on listening test scores.

Study 2 overview

In Study 2, a Gazepoint GP3 Eye Tracker (60 Hz data sampling rate, 0.5–1 degree of visual angle accuracy, 50–80cm operating distance, 1,920 × 1,080 display) was employed to record the eye movements of 15 test-takers while they were completing 24 computer-delivered listening test items. The test-takers were non-native speakers of English enrolled in advanced-level ESL courses at a public university in the Pacific region. The test items were adapted from the Michigan English Test (MET) and comprised three sets of listening items presented in a four-option multiple-choice format: discrete dialogue items (k = 10), dialogic listening items (k = 6), and monologic listening items (k = 8). All the listening prompts were administered as audio-only files and participants were allowed to take notes during the test using paper and pencil. Upon the completion of each listening item set, the test-takers participated in cued retrospective reporting, wherein the recordings of their eye movements were used to cue their reporting of the test-taking strategies used for answering each individual item. The participants were also instructed to elucidate the reasons for choosing a specific option in each test item, with all their responses being audio-recorded.

To determine specific types of test-taking strategies, including test management and test-wiseness strategies, the eye-movement recordings associated with each individual test item underwent a qualitative scanpath analysis via visual inspection (Ehmke and Wilson 2007), with a total of 360 scanpath data sets (i.e., 24 items × 15 participants). The analysis of each scanpath data set consisted of a written notation that described which elements of an item a test-taker had fixated on, the sequence of fixations, and the amount of time spent on the item, including any abnormalities or idiosyncrasies. The audio-recorded verbal data were transcribed and analysed qualitatively to identify the types of test-taking strategies reported by the participants and their reasons for choosing a particular answer. A strategy was marked as a test-wiseness strategy when a respondent provided no indication of knowing the answer to a specific question and reported strong doubts about the selected option being the correct answer. The results of the verbal data analysis were subsequently converged with the results of

the scanpath analysis to make final determinations of test-taking strategies. Next, given the study's goal of measuring the extent to which test-wiseness strategies affected test scores, a Wilcoxon Signed-Rank Test was used to determine whether there was a statistically significant difference between the observed scores from the test and the adjusted scores. The adjusted scores were calculated by subtracting the number of items that had been answered correctly using test-wiseness strategies from the observed scores.

The findings of this study evinced two main groups of test management strategies that the L2 learners resorted to when responding to the listening test items adopted from the MET: a) strategies related to the order of viewing or interacting with item elements, such as reading the question and response options while listening to the audio prompt or previewing the question and options before listening to the audio prompt (three types); and b) strategies used for interacting with the question and/or the response options and selecting the answer, such as re-reading the question and/or the response options several times before selecting the answer or selecting a response option while listening and skipping the rest of the audio prompt to move to the next question (nine types). Moreover, six types of test-wiseness strategies were also identified. Examples of test-wiseness strategies included selecting a response option by making a random guess, selecting a response option that looks different from other options, and selecting a response option by using background knowledge. Finally, the results of the Wilcoxon Signed-Rank Test indicated that observed scores (M = 20.2, SD = 2.83) were statistically significantly higher than adjusted scores (M = 18.07, SD = 3.86), $Z = -3.18$, $p < .01$, with a large effect size ($r = .58$).

Discussion and conclusion

An overarching goal of this chapter was to outline the need for more concerted efforts to complement product-oriented approaches with process-based approaches to L2 listening assessment validation via eye tracking. The two studies reported in this chapter attempted to demonstrate how eye-tracking technology could be used to gather the response process data (i.e., test-takers' viewing behaviour during the video-mediated L2 listening test in Study 1 and test-taking strategies used to answer listening test items in Study 2) that can serve as a source of the cognitive validity evidence in Weir's (2005) socio-cognitive framework.

The findings of Study 1 evinced that the test-takers watched content videos differently and for longer compared to context videos (as demonstrated by higher fixation rates and total dwell times), indicating that the test-takers found the visual information in content videos to be more informative for their listening comprehension than the visual information in context videos. Similar dwell rates, however, suggested that the participants deemed both

video types equally important. With regard to different aspects of the visual information in the videos, the test-takers considered lecture-related aspects to be more helpful for their listening comprehension and test performance than speaker-related aspects, most likely due to their semantic richness and relevance to the auditory information in the videos. In sum, information about response processes elicited via the combination of eye tracking and cued retrospective reporting in Study 1 indicated that different types of visuals had differing effects on test-takers' viewing behaviour and self-perceived listening effectiveness and test performance, thereby providing validity evidence for visuals being an important variable in the construct of L2 visual-inclusive listening comprehension.

Converging the eye-movement data with the verbal report data in Study 2 allowed for the detection of 12 types of test management strategies (divided into two groups) and six types of test-wiseness strategies used by the test-takers. The investigation of the test-takers' response processes in this study revealed that their use of test-wiseness strategies had a statistically significant effect on the observed test scores and contributed to construct-irrelevant variance. This cognitive validity evidence derived from the analysis of the response process data weakened the claims that could be made on the basis of the test-takers' scores in the study.

The two studies summarised in this chapter have several important implications for future work in this area. In line with previous research (e.g., Bax 2013, Brunfaut and McCray 2015), the two studies demonstrated that eye tracking can be a viable methodology for investigating cognitive validity and obtaining validity evidence based on response processes. When used in isolation, however, eye-tracking data representing test-takers' response processes may be hard to interpret (cf. Winke and Lim 2014) without additional elucidation obtained via cued retrospective reporting. In the two studies, triangulating eye-tracking data with verbal data appeared to be crucial to the findings, with an implication being that triangulation should be an indispensable component of validation research that uses eye tracking, especially when the response processes under investigation pertain to cognitive (i.e., why test-takers do what they do during the test) rather than behavioural aspects (i.e., what exactly test-takers do). This implication aligns with recommendations from other researchers who espouse eye-tracking triangulation (e.g., Godfroid and Schmidtke 2013, Lee and Winke 2017, Winke et al 2013).

Another implication is that eye-movement recordings have strong potential to facilitate test-takers' verbalisations when used as a stimulus for cued retrospective reporting. This was particularly evident in Study 2, where the participants' verbal descriptions of test-taking strategies used for answering each test item appeared to be noticeably detailed and descriptive as they were facilitated by the participants' viewing of their own eye-movement recordings. It should be acknowledged, however, that determining the extent

to which these verbal descriptions cued by the participants' eye-movement recordings were more detailed and precise compared to those taken from traditional, non-cued retrospective verbal report methods would require an experimental study. A separate study would also be needed to determine whether cueing low-proficiency L2 test-takers with the recordings of their eye movements may pose an additional cognitive load on them and make verbalising their cognitive processes a more cognitively demanding task than non-cued verbal reports.

The results of both studies also have implications for the design of L2 listening tests. In Study 1, a statistically significant difference was found between the two types of video, with content videos being more semantically important for L2 listeners (as suggested by higher fixation rates) and containing more meaningful information in the visual channel (as suggested by the higher percentage of total dwell time) than context videos. This finding indicates that visuals are essential for L2 listening comprehension and suggests that the ability to understand visual information should be part of the construct of L2 academic listening comprehension. Such results also highlight the importance of designing L2 listening test items that are capable of tapping into test-takers' comprehension of both the auditory and the visual input. With test-wiseness strategies playing a more pivotal role in the performance of low-proficiency test-takers compared to high-proficiency ones, Study 2 suggests a need to move away from the multiple-choice item format and, instead, adopt item formats that are less conducive to the use of strategies that introduce construct-irrelevant variance.

The potential of eye tracking as a methodology for gathering validity evidence based on response processes in L2 listening assessments is indisputably larger than what has been demonstrated by the two studies. Eye-tracking technology can be used to obtain procedural evidence based on the oculomotor data (e.g., the sequence of visual attention, such as the order in which test-takers view the options in a multiple-choice item or the number of times they re-read each option). Analysing eye-tracking metrics such as response time data for each item in a listening test can reveal test-takers' use of time-allotment strategies, including the use of particularly prolonged or particularly short response times. In addition, the scanpath analysis of eye-movement data can expose aberrant response patterns (e.g., responding to a multiple-choice item before hearing the prompt) and evince certain abnormalities or idiosyncrasies in individual test-taking behaviours, such as random guessing, rapid guessing, response latencies, and response changing.

Eye tracking can also be instrumental for detecting the effect of various factors – such as task design, test design, and test-takers' emotional states – on response processes. Given that the viewers' visual attention can be driven to a significant extent by their emotions (Conklin et al 2018), the potential of eye tracking for gathering validity evidence based on response processes

can be enhanced by complementing the eye gaze data with information about test-takers' affective or emotional states (such as the level of anxiety, frustration, interest, boredom, or enjoyment) elicited via verbal report methods. Research questions of eye-tracking studies in the area of L2 listening assessment can enquire into how test-takers' response processes and test performances are affected by different ways of presenting the test content (such as displaying the video prompt and the questions on the same screen vs. playing the video prompt first before showing the test items). Furthermore, harnessing this technology can provide evidence of the differences among test-takers with different L2 listening comprehension abilities regarding their use of task-relevant auditory and/or visual information while completing L2 listening assessment tasks (e.g., whether those test-takers who direct their gaze more quickly to the test item when they hear the relevant information in the auditory stimulus perform better than those test-takers whose reaction time is slower).

It is important to recognise that eye tracking is not a panacea for tackling all the problems or alleviating all the concerns in process-oriented L2 listening assessment validation. For instance, eye tracking can be less accurate and useful in computer-assisted language testing contexts that allow for paper-based note-taking because those test-takers who are committed note-takers will not be looking at the screen and, consequently, their eye movements will not be recorded at the time of note-taking. However, although allowing paper-based note-taking in both studies had a negative effect on the quality and continuity of eye-tracking data recording, the extent of this effect was not substantial. Specifically, in Study 1 participants concentrated mostly on watching the video prompts, which resulted in the low-level note-taking activity. In Study 2, the effect of note-taking was even less of an issue since the study focused primarily on the visual scanpath analysis of the eye-tracking data representing the participants' responses to multiple-choice questions rather than the data representing their listening to the prompts and note-taking. When designing similar eye-tracking studies in the future, researchers need to take into consideration their research questions, the types of eye-tracking data they need to gather, and the types of eye-tracking data analyses they plan to carry out when deciding whether to allow paper-based note-taking or not.

In addition, eye tracking, as mentioned earlier, does not provide direct evidence related to cognition; information about cognitive processes underlying test-takers' response processes can only be inferred indirectly from the eye-movement data through the eye-mind hypothesis (Just and Carpenter 1980), oftentimes by triangulating them with other types of response process data such as verbal report data. Consequently, when analysing eye-tracking data, researchers would generally have limited scope to tease apart test-takers' cognitive processes that occur simultaneously (cf. Bax 2013).

Given the limitations of the eye-mind hypothesis, certain incongruencies between test-takers' oculomotor behaviour and mental processes, such as gazing extensively and aimlessly at one item in the visual field while pondering something unrelated to that item, would also be impossible to detect solely from tracking test-takers' eye movements.

Another important issue that is sometimes overlooked is the limitations of eye-tracking technology itself. As the market is increasingly saturated with eye-tracking hardware and software of varying price and quality, the risks of studies fraught with serious methodological flaws also rise due to variations in the quality of eye-tracking data and the interpretations that can be made based on their analyses. The two studies reported in this chapter, for instance, used two different entry-level eye-tracking systems with low data sampling rates. With the overall rate of eye detection in Study 1 being 94%, the quality of the eye-tracking data was relatively high and was not deemed to have a significantly detrimental effect on the accuracy of the findings. In Study 2, the rate of eye detection was not measured by the software program; however, the researcher verified the accuracy of calibration and monitored eye tracking during each data collection session to ensure the integrity of the recorded data was acceptable for the visual scanpath analysis. Overall, while the eye-tracking data recorded by these systems were not highly precise, the data quality was adequate enough for answering the RQs posed in the two studies. However, these entry-level systems would be inadequate, for example, for a psycholinguistic type of research or for a study of L2 reading that entails analyses of the data retrieved from multiple, tightly clustered AOIs. The design of such a study would be questionable at best and may result in bogus findings, misguided claims, and invalid conclusions or, even worse, get cited and adopted by other unwitting researchers. To avoid poorly designed eye-tracking studies, language testers and researchers interested in investigating this method should first get acquainted with the basics of eye-tracking research and common misconceptions rather than attempt to 'play around' and run 'exploratory' studies. A good starting point for novice researchers interested in employing eye tracking for researching listening, for example, is Chapter 5 in the guide for applied linguistics eye-tracking research by Conklin et al (2018).

Given the pace and scope of technological advancements, as well as the progress in gathering user analytics, one can expect that eye tracking may soon become mainstream technology integrated into personal computers, thereby creating new opportunities for language testers to collect eye-tracking data together with test scores during high-stakes language testing. Such recordings of test-takers' eye movements on a large scale could become an integral part of any computer-based language test and be used as an additional source of cognitive validity evidence, thereby increasing the test

developers' confidence in the test scores they report and the interpretations they can make on the basis of the test scores with respect to test-takers' L2 proficiency.

Acknowledgement

Research reported in this chapter was supported by the Small Grant for Doctoral Research in Second Language Assessment from Educational Testing Service (ETS), the Assessment Research Grant from the British Council, and the Spaan Research Grant Program 2016 from Michigan Language Assessment (MLA).

References

American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME] and Joint Committee on Standards for Educational and Psychological Testing (2014) *Standards for Educational and Psychological Testing*, Washington, D.C.: American Educational Research Association.

Anderson, N J, Bachman, L F, Perkins, K and Cohen, A D (1991) An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources, *Language Testing* 8 (1), 41–66.

Aryadoust, V (2019a) An integrated cognitive theory of comprehension, *International Journal of Listening* 33, 77–100.

Aryadoust, V (2019b) Dynamics of item reading and answer changing in two hearings in a computerized while-listening performance test: An eye-tracking study, *Computer Assisted Language Learning* 32, 1–28.

Bachman, L F (2005) Building and supporting a case for test use, *Language Assessment Quarterly* 2 (1), 1–34.

Bachman, L F and Palmer, A S (2010) *Language Assessment in Practice*, Oxford: Oxford University Press.

Batty, A O (2015) A comparison of video- and audio-mediated listening tests with many-facet Rasch modeling and differential distractor functioning, *Language Testing* 32 (1), 3–20.

Batty, A O (2017) *The impact of visual cues on item response in video-mediated tests of foreign language listening comprehension*, unpublished doctoral dissertation, Lancaster University, available online: eprints.lancs.ac.uk/id/eprint/84467/4/batty_thesis_news.pdf

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

Bax, S and Chan, S H C (2016) *Researching the Cognitive Validity of GEPT High-Intermediate and Advanced Reading: An Eye Tracking and Stimulated Recall Study*, Tapei: The Language Training and Testing Center (LTTC), available online: oro.open.ac.uk/47502/

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

Bejar, I, Douglas, D, Jamieson, J, Nissan, S and Turner, J (2000) *TOEFL 2000 Listening Framework: A Working Paper*, Princeton: Educational Testing Service.

Bodie, G D, Janusik, L A and Välikoski, T-R (2008) *Priorities of listening research: Four interrelated initiatives*, white paper sponsored by the Research Committee of the International Listening Association, available online: dokumen.tips/documents/white-paper-priorities-of-listening-research-31808-of-listening-research-four.html?page=1

Bowles, M A (2010) *The Think-aloud Controversy in Second Language Research*, New York: Routledge.

Brunfaut, T (2016) *Looking into Reading II: A Follow-up Study on Test-takers' Cognitive Processes While Completing Aptis B1 Reading Tasks*, British Council Validation Series, Volume VS/2016/001, London: The British Council.

Brunfaut, T and McCray, G (2015) *Looking into test-takers' cognitive processes whilst completing reading tasks: A mixed-method eye-tracking and stimulated recall study*, ARAGs Research Reports Online, Volume AR/2015/001, London: British Council, available online: www.britishcouncil.org/sites/default/files/brunfaut-and-mccray-report_final.pdf

Buck, G (2001) *Assessing Listening*, Cambridge: Cambridge University Press.

Chapelle, C A (2011) Validity argument for language assessment: The framework is simple…, *Language Testing* 29 (1), 19–27.

Chapelle, C A, Enright, M K and Jamieson, J M (Eds) (2008) *Building a Validity Argument for the Test of English as a Foreign Language*, New York: Routledge.

Cohen, A D (2007) The coming of age for research on test-taking strategies, in Fox, J, Weshe, M, Bayliss, D, Cheng, L, Turner, C and Doe, C (Eds) *Language Testing Reconsidered*, Ottawa: Ottawa University Press, 89–111.

Cohen, A D (2014) Using test-wiseness strategy research in task development, in Kunnan, A J (Ed) *The Companion to Language Assessment*, Malden: Wiley-Blackwell, 893–905.

Cohen, A D and Upton, T A (2007) 'I want to go back to the text': Response strategies on the reading subtest of the new TOEFL®, *Language Testing* 24 (2), 209–250.

Coniam, D (2001) The use of audio or video comprehension as an assessment instrument in the certification of English language teachers: A case study, *System* 29, 1–14.

Conklin, K and Pellicer-Sánchez, A (2016) Using eye-tracking in applied linguistics and second language research, *Second Language Research* 32 (3), 453–467.

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

Cubilo, J and Winke, P M (2013) Redefining the L2 listening construct within an integrated writing task: Considering the impacts of visual-cue interpretation and note-taking, *Language Assessment Quarterly* 10 (4), 371–397.

Duchowski, A (2007) *Eye-tracking Methodology: Theory and Practice* (Second edition), London: Springer-Verlag.

Ehmke, C and Wilson, C (2007) *Identifying web usability problems from eye-tracking data*, paper presented at the British HCI conference 2007, University of Lancaster, September.

Ercikan, K and Pellegrino, J W (Eds) (2017) *Validation of Score Meaning for the Next Generation of Assessments*, New York: Routledge.

Field, J (2008) *Listening in the Language Classroom*, Cambridge: Cambridge University Press.

Field, J (2013) Cognitive validity, in Geranpayeh, A and Taylor, L (Eds) *Examining Listening: Research and Practice in Assessing Second Language*

Listening, Studies in Language Testing Volume 35, Cambridge: UCLES/ Cambridge University Press, 77–151.

Flowerdew, J and Miller, L (2010) Listening in a second language, in Wolvin, A D (Ed) *Listening and Human Communication in the 21st Century*, Oxford: Wiley-Blackwell, 158–177.

Ginther, A (2002) Context and content visuals and performance on listening comprehension stimuli, *Language Testing* 19 (2), 133–167.

Godfroid, A and Schmidtke, J (2013) What do eye movements tell us about awareness? A triangulation of eye-movement data, verbal reports, and vocabulary learning scores, in Bergsleithner, J M, Frota, S N and Yoshioka, J K (Eds) *Noticing and Second Language Acquisition: Studies in Honor of Richard Schmidt*, Honolulu: University of Hawai'i at Manoa, National Foreign Language Resource Center, 183–205.

Godfroid, A and Winke, P M (2015) Investigating implicit and explicit processing using L2 learners' eye-movement data, in Rebuschat, P (Ed) *Implicit and Explicit Learning of Languages*, Amsterdam: John Benjamins, 325–348.

Godfroid, A, Winke, P M and Gass, S (Eds) (2013) Thematic issue on eye-tracking in second language acquisition research, *Studies in Second Language Acquisition* 35 (2).

Goh, C C M and Aryadoust, V (2016) Learner listening: New insights and directions from empirical studies, *The International Journal of Listening* 30, 1–7.

Gruba, P (1993) A comparison study of audio and video in language testing, *JALT Journal* 15 (1), 85–88.

Harding, L (2012) Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective, *Language Testing* 29 (2), 163–180.

Holmqvist, K, Nyström, M, Andersson, R, Dewhurst, R, Jarodzka, H and Van de Weijer, J (2011) *Eye-tracking: A Comprehensive Guide to Methods and Measures*, Oxford: Oxford University Press.

Holzknecht, F, Eberharter, K, Kremmel, B, Zehentner, M, McCray, G, Konrad, E and Spöttl, C (2017) *Looking into Listening: Using Eye-tracking to Establish the Cognitive Validity of the Aptis Listening Test*, ARAGs Research Reports Online, Volume AR-G/2017/3, available online: www.britishcouncil.org/exam/aptis/research/publications/arags/looking-listening-using-eye-tracking

Hubley, A M and Zumbo, B D (2017) Response processes in the context of validity: Setting the stage, in Zumbo, B D and Hubley, A M (Eds) *Understanding and Investigating Response Processes in Validation Research*, New York: Springer, 1–12.

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

Kane, M T (1992) An argument-based approach to validity, *Psychological Bulletin* 112 (3), 527–535.

Kane, M T (2006) Validation, in Brennan, R (Ed) *Educational Measurement* (Fourth edition), Westport: Praeger, 17–64.

Kane, M T (2010) Validity and fairness, *Language Testing* 27 (2), 177–182.

Kane, M T and Mislevy, R J (2017) Validating score interpretations based on response processes, in Ercikan, K and Pellegrino, J W (Eds) *Validation of Score Meaning for the Next Generation of Assessments*, New York: Routledge, 11–24.

Kashkouli, Z and Barati, H (2013) Type of test-taking strategies and task-based reading assessment: A case in Iranian EFL learners, *Procedia – Social and Behavioral Sciences* 70, 1,580–1,589.

Launeanu, M and Hubley, A M (2017a) A model building approach to examining response processes as a source of validity evidence for self-report items and measures, in Zumbo, B D and Hubley, A M (Eds) *Understanding and Investigating Response Processes in Validation Research*, New York: Springer, 115–136.

Launeanu, M and Hubley, A M (2017b) Some observations on response processes research and its future theoretical and methodological directions, in Zumbo, B D and Hubley, A M (Eds) *Understanding and Investigating Response Processes in Validation Research*, New York: Springer, 93–113.

Lee, H and Winke, P M (2012) The difference among three-, four-, and five-option-item formats in the context of a high-stakes English-language listening test, *Language Testing* 30 (1), 99–123.

Lee, S and Winke, P M (2017) Young learners' response processes when taking computerized tasks for speaking assessment, *Language Testing* 35 (2), 239–269.

Londe, Z C (2009) The effects of video media in English as a second language listening comprehension tests, *Issues in Applied Linguistics* 17 (1), 41–50.

Lynch, T (2011) Academic listening in the 21st century: Reviewing a decade of research, *Journal of English for Academic Purposes* 10, 79–88.

McCray, G and Brunfaut, T (2018) Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking, *Language Testing* 35 (1), 51–73.

Messick, S (1995) Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning, *American Psychologist* 50 (9), 741–749.

Mislevy, R J (2009) *Validity From the Perspective of Model-based Reasoning*, CRESST Report 752, Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, available online: cresst.org/wp-content/uploads/R752.pdf

Morley, J (2001) Aural comprehension instruction: Principles and practices, in Celce-Murcia, M (Ed) *Teaching English as a second or foreign language* (Third edition), Boston: Heinle and Heinle, 69–85.

Ockey, G J (2007) Construct implications of including still image or video in computer-based listening tests, *Language Testing* 24 (4), 517–537.

Oranje, A, Gorin, J, Jia, Y and Kerr, D (2017) Collecting, analyzing, and interpreting response time, eye-tracking, and log data, in Ercikan, K and Pellegrino, J W (Eds) *Validation of Score Meaning for the Next Generation of Assessments*, New York: Routledge, 39–51.

Padilla, J-L and Benítez, I (2014) Validity evidence based on response processes, *Psicothema* 26 (1), 136–144.

Phakiti, A (2003) A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance, *Language Testing* 20 (1), 26–56.

Plakans, L (2009) Discourse synthesis in integrated second language writing assessment, *Language Testing* 26 (4), 561–587.

Pusey, K and Lenz, K (2014) Investigating the interaction of visual input, working memory, and listening comprehension, *Language Education in Asia* 5, 66–80.

Rayner, K (2009) Eye movements and attention in reading, scene perception, and visual search, *The Quarterly Journal of Experimental Psychology* 62 (8), 1,457–1,506.

Révész, A, Michel, M and Lee, M (2017) *Investigating IELTS Academic Writing Task 2: Relationships Between Cognitive Writing Processes, Text Quality, and Working Memory*, IELTS Research Reports Online Series 2017/3, available online: search.informit.org/doi/book/10.3316/informit.840774005843694

Robinson, P, Mackey, A, Gass, S M and Schmidt, R (2012) Attention and awareness in second language acquisition, in Gass, S M and Mackey, A (Eds) *The Routledge Handbook of Second Language Acquisition*, New York: Routledge, 247–267.

Rost, M (2002) *Teaching and Researching Listening*, London: Longman.

Rost, M (2014) Listening in a multilingual world: The challenges of second language (L2) listening, *International Journal of Listening* 28, 131–148.

Rost, M and Wilson, J J (2013) *Active Listening*, New York: Routledge.

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

Schmitt, N, Ng, J W C and Garras, J (2011) The Word Associates Format: Validation evidence, *Language Testing* 28 (1), 105–126.

Seo, D, Taherbhai, H and Frantz, R (2016) Psychometric evaluation and discussions of English language learners' listening comprehension, *International Journal of Listening* 30, 47–66.

Shin, D (1998) Using videotaped lectures for testing academic listening proficiency, *International Journal of Listening* 12 (1), 57–80.

Sueyoshi, A and Hardison, D (2005) The role of gestures and facial cues in second language listening comprehension, *Language Learning* 55 (4), 661–699.

Suvorov, R (2009) Context visuals in L2 listening tests: The effects of photographs and video vs. audio-only format, in Chapelle, C A, Jun, H G and Katz, I (Eds) *Developing and Evaluating Language Learning Materials*, Ames: Iowa State University, 53–68.

Suvorov, R (2013) *Interacting with visuals in L2 listening tests: An eye-tracking study*, unpublished doctoral dissertation, Iowa State University.

Suvorov, R (2015) The use of eye-tracking in research on video-based second language (L2) listening assessment: A comparison of context videos and content videos, *Language Testing* 32 (4), 463–483.

Suvorov, R (2018a) *Investigating Test-taking Strategies During the Completion of Computer-delivered Items From Michigan English Test (MET): Evidence From Eye-tracking and Cued Retrospective Reporting*, Cambridge Michigan Language Assessment (CaMLA) Working Papers 2018-02, available online: michiganassessment.org/wp-content/uploads/2020/02/20.02.pdf.Res_. MichiganEnglishTestMET-EvidencefromeyeTrackingandCuedRetrospective Reporting.pdf

Suvorov, R (2018b) Test-takers' use of visual information in an L2 video-mediated listening test: Evidence from cued retrospective reporting, in Ockey, G and Wagner, E (Eds) *Assessing L2 Listening: Moving Towards Authenticity*, Amsterdam: John Benjamins, 145–160.

Taylor, L (2013) Introduction, in Geranpayeh, A and Taylor, L (Eds) *Examining Listening: Research and Practice in Assessing Second Language Listening*, Studies in Language Testing Volume 35, Cambridge: UCLES/Cambridge University Press, 1–35.

Taylor, L and Geranpayeh, A (2011) Assessing listening for academic purposes: Defining and operationalizing the test construct, *Journal of English for Academic Purposes* 10, 89–101.

van der Linden, W J (2009) Conceptual issues in response-time modeling, *Journal of Educational Measurement* 46 (3), 247–272.
Van Gog, T, Paas, F, Van Merriënboer, J J G and Witte, P (2005) Uncovering the problem-solving process: Cued retrospective reporting versus concurrent and retrospective reporting, *Journal of Experimental Psychology: Applied* 11 (4), 237–244.
Vandergrift, L (2010) Researching listening, in Paltridge, B and Phakiti, A (Eds) *Continuum Companion to Research Methods in Applied Linguistics*, London: Continuum International Publishing Group, 160–173.
Vandergrift, L (2012) Teaching listening, in Chapelle, C A (Ed) *The Encyclopedia of Applied Linguistics*, available online: doi.org/10.1002/9781405198431.wbeal1169
Wagner, E (2007) Are they watching? Test-taker viewing behavior during an L2 video listening test, *Language Learning & Technology* 11 (1), 67–86.
Wagner, E (2010a) Test-takers' interaction with an L2 video listening test, *System* 38, 280–291.
Wagner, E (2010b) The effect of the use of video texts on ESL listening test-taker performance, *Language Testing* 27 (4), 493–513.
Wagner, E (2013) An investigation of how the channel of input and access to test questions affect L2 listening test performance, *Language Assessment Quarterly* 10 (2), 178–195.
Weir, C (2005) *Language Testing and Validation: An Evidence-based Approach*, Basingstoke: Palgrave Macmillan.
Winke, P M and Lim, H (2014) *The effects of testwiseness and test-taking anxiety on L2 listening test performance: A visual (eye-tracking) and attentional investigation*, IELTS Research Reports Online Series 3, IELTS Partners: British Council/IDP/Cambridge English.
Winke, P M and Lim, H (2015) ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study, *Assessing Writing* 25 (2), 37–53.
Winke, P M, Gass, S and Sydorenko, T (2013) Factors influencing the reading of captions by foreign language learners: An eye-tracking study, *Modern Language Journal* 97 (1), 254–275.
Winke, P M, Lee, S, Ahn, J I, Choi, I, Cui, Y and Yoon, H-J (2018) The cognitive validity of child English language tests: What young language learners and their native-speaking peers can reveal, *TESOL Quarterly* 52 (2), 274–303.
Wu, A D and Stone, J E (2015) Validation through understanding test-taking strategies: An illustration with the CELPIP-General Reading Pilot Test using structural equation modeling, *Journal of Psychoeducational Assessment* 34 (4), 362–379.
Wu, A D and Zumbo, B D (2017) Understanding test-taking strategies for a reading comprehension test via latent variable regression with Pratt's importance measures, in Zumbo, B D and Hubley, A M (Eds) *Understanding and Investigating Response Processes in Validation Research*, New York: Springer, 305–319.
Yamashita, J (2003) Processes of taking a gap-filling test: Comparison of skilled and less skilled EFL readers, *Language Testing* 20 (3), 267–293.
Zumbo, B D and Hubley, A M (Eds) (2017) *Understanding and Investigating Response Processes in Validation Research*, New York: Springer.

4 A comparative study on audio-only and video-based listening tests: The impact of visual input

Suh Keong Kwon
Chinju National University of Education, Republic of Korea

Abstract

This study examines the extent to which test-takers' listening test performances and viewing patterns were affected when a video prompt was introduced to an L2 listening comprehension test. A total of 117 Korean learners of English were divided into two groups, each given the same listening test in two different modes: an audio-only mode and a video mode. The two groups were compared on their test scores and eye movements. For eye-movement analysis, three different fixation measures were analysed to investigate the extent to which test-takers viewed each Area of Interest (AOI) during the test.

Findings indicated that the video group performed significantly better than the audio-only group, but this difference was not substantial. The results of the eye-movement analysis showed that test-takers in the audio-only group looked at the questions and options significantly longer and more frequently than those in the video group. This difference in eye movement was likely caused by the additional visual input the video group received. It was found that the video group viewed the speaker(s) and the visual aid (PowerPoint slides in the video) for more than half of the total test time. These findings suggest that the presence of visual input in the listening test caused substantial changes in test-takers' viewing patterns, though it did not have a strong effect on their test performances. The study concludes that presenting visual input in L2 listening tests can be a desirable approach to improving their validity.

Introduction

Video is increasingly used for language learning but its use in language tests remains tentative. Since most real-life listening activities are enhanced by visuals, researchers in language testing have been keen to investigate whether

adding visual input to a listening test affects listeners' cognitive processes and test performances. Previous studies have mostly explored the effect of visual input in a listening test, mainly by comparing test performances between audio-only and video modes of L2 listening comprehension tests (Coniam 2001, Cubilo and Winke 2013, Ginther 2002, Suvorov 2009, 2015, Wagner 2007, 2008). These studies yielded controversial findings – whereas some studies found a positive effect of video on listening performance (Cubilo and Winke 2013, Wagner 2007, 2008, 2010, 2013), the others found either no effect (Coniam 2001, Gruba 1993) or even a negative effect (Ginther 2002, Suvorov 2009). Few studies have explored how and what specific visual cues are used by test-takers in listening comprehension, mainly due to methodological limitations. With the introduction of eye-tracking technology, language testing researchers can now identify the extent to which test-takers view specific visual cues during an L2 listening test (Batty 2017, Suvorov 2015). However, our knowledge of how test-takers view different kinds of visual input and the effect of viewing patterns on their test performances is still very limited. In addition, little research has examined this effect when the question and the answer choices are presented together in one test screen.

Literature review

Visual input in listening assessment

A large body of research has suggested that providing visual input in a listening comprehension test aids listeners in comprehending aural input and thus improves test-takers' listening performances (Cubilo and Winke 2013, Hernandez 2005, Suvorov 2015, Wagner 2007, 2008, 2010, 2013). However, a few other studies have found that adding visuals has no effect or even a debilitating effect on test-takers' performances (Buck 2001, Coniam 2001, Ginther 2002, Gruba 1993, Londe 2009, Suvorov 2013). Bejar, Douglas, Jamieson, Nissan and Turner (2000) state that the purpose of presenting video as a listening prompt is to enhance face validity and authenticity but, simultaneously, it could distract test-takers. Batty (2015) argues that a video listening test is more authentic with greater face validity but we know little about the specific visual cues that draw test-takers' attention or present supportive information. In response to the call for filling this research gap, Suvorov (2018) investigated listeners' varying degrees of comprehension when encountering different types of visual information in an academic listening test and revealed that context visuals, which contain speaker-related visual cues (e.g., the speaker's face), were less helpful than the content visuals (e.g., images showing semantically congruent ideas of the lecture). He speculated that the helpfulness of content visuals to comprehension could be attributed to the semantic values they contained.

Two studies by Batty (2015, 2018) also investigated the effect of visual input on listening performance by different item types. The 2018 study, for instance, compared the effect of video between implicit (finding global/main ideas) and explicit items (finding specific ideas), and found that implicit items were significantly easier in the video mode of the test. Taken together, such mixed findings suggest that a consensus has not been established as to whether it is appropriate to include visual input in a listening test.

In addition, a small number of studies investigated how test-takers engage with the visual input while listening. Ockey (2007) and Wagner (2007, 2010) used video cameras to record test-takers' attention and measure the amount of time they attend to the screen. However, it remains an open question as to what visual information test-takers pay attention to, and how their interaction with the video stimuli may differ according to types of listening task. Ockey (2007) also employed a retrospective verbal protocol procedure to investigate whether the visual input was helpful or distracting to listening comprehension. However, the findings did not sufficiently explain the rationales behind test-takers' viewing patterns and test-taking strategies due to the limited capacity of participants' memory. To overcome such a methodological limitation, Suvorov (2015) employed an eye-tracking method to measure L2 listening test-takers' gaze patterns and found that they interacted extensively with visual input during the listening test. However, Suvorov's (2015) study is limited in that it mainly focused on the effect of context and content visuals on viewing behaviours and listening performance, and that the listening prompts he used in the study were restricted to academic lectures. He suggested future studies examining the effect of more specific visual input beyond the context and content distinctions (Suvorov 2015). Winke and Lim (2014) also employed an eye-tracking method to examine test-takers' processing of visual and textual information and the extent to which test-wiseness strategies (i.e., guessing an answer by manipulating the test format) and test anxiety had an effect on test-takers' listening performance. However, they did not investigate what drew test-takers' attention and whether such viewing behaviour may improve their test performance.

On the other hand, a small number of eye-tracking studies on L2 learning have investigated what draws an L2 learner's attention when visual input is provided along with aural input. Mestres and Pellicer Sánchez (2019) found that test-takers spent a larger proportion of time viewing the visual input than reading subtitles (text) when watching a video. Additionally, they found that learners spent a longer proportion of time reading the text than looking at the images when the prompt is akin to an audio book (audio input and text). Muñoz (2017) found significant differences in mean fixations on the subtitles of cartoon videos between beginner and intermediate-level L2

learners. The study concludes that beginners skipped subtitles less frequently than more proficient learners.

Although previous works have examined the effect of visual input on listening comprehension, little attention has been paid to specific cognitive processes involved in decoding visual input in a video-mediated listening test. The current study aimed to fill several research gaps in the literature. Firstly, studies on listening processes have mainly focused on types of the visual input such as context versus content visuals (Suvorov 2015) and still images versus videos (Ockey 2007), and listener behaviours, such as the amount of time spent on watching the video (Wagner 2010), and note-taking (Cubilo and Winke 2013). Investigations on more specific features in video-enhanced listening tests, such as PowerPoint slides or stem-and-answer choices, are required as these are commonly used in conventional audio-only listening tests seen in the Internet-based Test of English as a Foreign Language (TOEFL iBT) and IELTS. Also, little is known about how test-takers interact with the visual input such as images of the speakers or the slides. Secondly, many previous studies compared test-takers' performances in audio and audio-visual conditions but only a few have investigated individual differences in viewing behaviours (Suvorov 2015, Winke and Lim 2014, Winke, Gass and Sydorenko 2013). Lastly, previous studies have not extensively investigated the effect of visuals on various listening situations as most focused on testing English for Academic Purposes (EAP) in a higher education context. The listening materials used in this present study include a wide range of listening situations (both dialogues and lectures) that simulate real-world listening activities to improve the context validity of the listening tests (Bachman and Palmer 1996, Buck 2001). Also, the visual input used in this study contains both content-related and context-related information as Suvorov (2015) argues that there is no content visual without any context. Other accessibility tools (i.e., timing device, highlighter, etc.) that are often built into conventional computer-based tests (CBTs) are not considered in this study.

The socio-cognitive approach to test validation

This study deems the listening construct as interactions between cognitive and contextual factors, or 'internal mental processes' and 'external contextual features' (Taylor 2013:25). Previous approaches towards the interaction between cognitive and context validity mainly focused on the investigation of test design factors such as the number of times listening input is heard, speech rate, various accents, and the number of speakers. However, the important role of visual input in a listening test has only been acknowledged by a few studies (Geranpayeh and Taylor (Eds) 2013, Li 2013). Presentation of visual input in a listening test is also closely involved with test-takers' internal

cognitive processes of decoding aural and visual input, and the contextual features that create proximity to real-world listening situations. The ability to process audio-visual input is considered an essential listening skill required in the target language use (TLU) domain, and is also illustrated as a defining characteristic of the six levels of the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001).

According to Weir's (2005) socio-cognitive framework, internal mental processing is often shaped by various external contextual factors. In his model, presentation of multimodal input is introduced as one of the task-setting components under context validity, and decoding visual input is also listed as one of the cognitive validity elements. Context validity mainly deals with the degree to which a given task in a test is 'representative' of the target context where the test is predicted to be a sample (Weir 2005:19). In language testing, we are interested in making interpretations about test-takers' language ability in the TLU domain (Bachman and Palmer 2010). Therefore, test-takers are expected to perform tasks that closely resemble real-life listening situations. In line with this, Vandergrift and Goh (2012) state that tasks that reflect real-life purposes of listening are 'authentic', and an authentic test is established by a relationship between the task and the listener. Elliott and Wilson (2013) also suggest that visual information is an important 'facet of context' in interpersonal interactions as non-verbal signals carry a significant 'social meaning' (2013:189). Based on these theories, we can assume that the visual input in a listening test can be one of the positive external factors that offer test-takers advantages in terms of both internal cognitive processes as well as produced output. This assumption may be checked by investigating whether test-takers utilise different cognitive processes when completing video-mediated listening tasks and traditional audio-only listening tasks.

Cognitive processes involved in listening comprehension with visual input

To measure listeners' cognitive processes, we need to develop tasks that elicit mental processes adequately representative of what listeners would employ in a TLU domain (Field 2013). In other words, simply measuring linguistic ability alone is not enough to make valid decisions from the test. Rather, how test-takers manipulate the given tasks and texts within the context to produce answers should also be taken into account in the test design. While a large number of studies have looked into the effect of visual input on listening process and performance, only a handful of them investigated test-takers' viewing behaviours and their engagement with visual cues (e.g., Batty 2020, Cubilo and Winke 2013, Lesnov 2018, Ockey 2007, Suvorov 2015, 2018, Wagner 2007, 2008, 2010, 2013).

To explain the cognitive processes involved in a listening test, Bejar et al (2000) modelled listening comprehension by splitting it into 'listening' and 'response' stages. This model, as shown in Figure 1, describes three different types of knowledge that test-takers draw on to formulate responses in a listening test, namely situational knowledge, linguistic knowledge, and background knowledge. According to this model, listeners receive both aural and visual input when accessing the three types of knowledge, and then create a set of propositions to understand what they have heard (Bejar et al 2000). Among these three types of knowledge, situational knowledge is closely associated with the visual input because it deals with the role of context of the input. Listeners often see the background scene and the person they are listening to, which provide clues for understanding the meaning of the aural input. A few studies suggest that this complex process of listening may be affected by listeners' access to their background knowledge of the situation (Bejar et al 2000, Cubilo and Winke 2013, Gruba 1993).

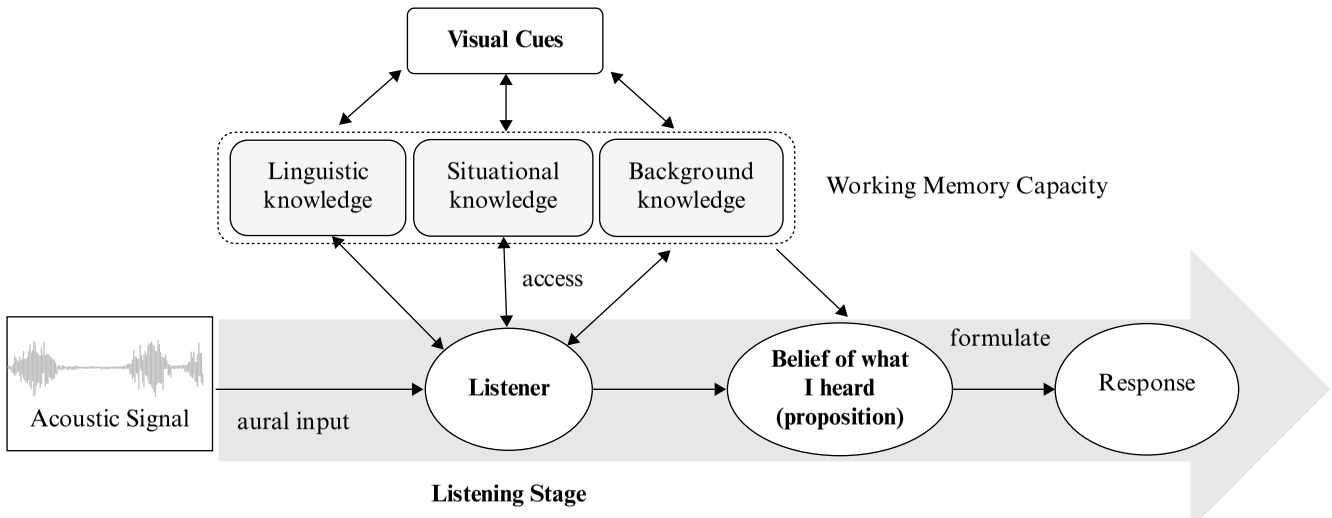
Recently, eye-tracking research has received attention as it provides further insight into test-takers' cognitive processes while completing a listening test. A number of existing eye-tracking studies have shed light on how test-takers interact with visual input and how different types of visual input affect test-takers' listening processes (Batty 2020, Suvorov 2015). However, research on test-takers' viewing behaviours in a listening test is still in its infancy and issues such as how they interact with PowerPoint slides or stem-and-answer choices are unexplored. Test-takers' eye movements would be very useful data to understand such test-taking processes since it provides information about the efforts that test-takers spend on comprehending textual and visual input and also reflects where their attention is directed to (Conklin, Pellicer-Sánchez and Carrol 2018).

In summary, the existing body of research on video-mediated listening comprehension falls short in explaining how and to what extent test-takers utilise the visual input in their listening processes. To narrow the research gap identified, this study investigates what specific visual cues are viewed by test-takers while sitting a video-mediated listening test and how these viewing behaviours might affect their test-taking processes as compared to sitting an audio-only listening test. With this in mind, this study addresses the following two research questions (RQs):

RQ1: To what extent does listening performance differ between audio-only and video-based listening conditions?

RQ2: In what way does the presence of visual input in a listening comprehension test change test-takers' viewing behaviours?

Figure 1 Model of listening and response stages (adapted and expanded from Bejar et al 2000)



Methodology

Participants

The participants of this study were 117 pupils enrolled in five different secondary schools in the Republic of Korea (Korea, hereafter). Of these participants, 57 were allocated to the control group (audio-only) and 60 were allocated to the experimental group (video) randomly. The recent National English Listening Test (NELT)[1] scores of both groups were compared to ensure the two groups were generally equivalent in listening ability. The gender distribution and mean age were almost identical between the two groups: the audio-only group had 27 male and 30 female participants with a mean age of 16.93; the video group had 30 male and 30 female participants with a mean age of 16.65. In the eye-tracking test, all but seven pupils (three video, four audio-only), who failed to calibrate with the eye-tracking device, participated.

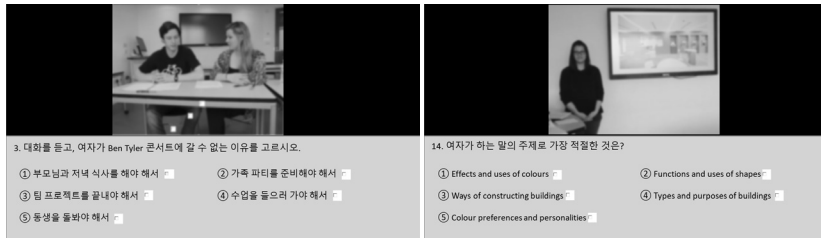
Video-mediated listening items

Fifteen retired items from the College Scholastic Ability Test (CSAT) and NELT were selected and adapted to construct the video-enhanced listening test for this study, considering both their relevance to the TLU domain and feasibility. The CSAT is a high-stakes national exam administered for university admission purposes, and NELT is a national English listening test for diagnosing students' English listening proficiency. Of these 15 items, eight of them were dialogues produced by two speakers of opposite gender, and seven of them were short lectures given by a single speaker. All items were five-option multiple-choice questions each with one correct answer. The instructions and stems were presented in Korean, the participants' L1. The options were also presented in Korean, except for inferencing items (choosing the best response following a conversation) which were given in English. The participants were familiar with these item types because they had taken similar listening tests as school exams and in CSAT preparation lessons.

The video prompts were developed based on prompt scripts identical to those used in the original audio-only tests. In the dialogue prompt, two speakers had a conversation while showing gestures and facial expressions. In the lecture items, one speaker gave a lecture showing hand gestures and

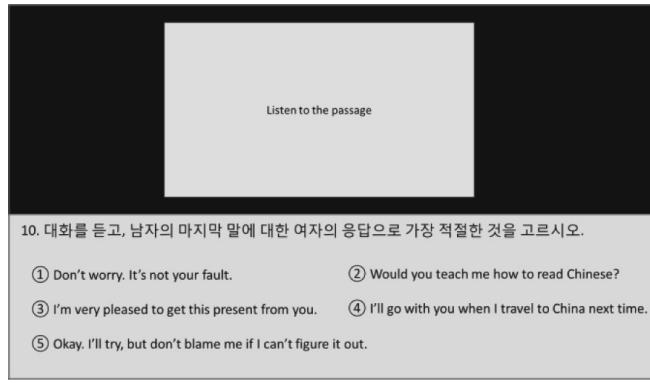
1 NELT is a national English listening test that is developed and administered by the Korea Institute for Curriculum and Evaluation. NELT is administered nationally twice a year for diagnostic and achievement testing purposes and the scores are usually used as in-school English grades in Korea.

Figure 2 Screenshot of the video-mediated listening test (left: dialogue, right: lecture)



facial expressions. There was also a PowerPoint slide present on the computer screen next to the speaker which presented images or text (words and phrases) that are relevant to the content of the lecture (see Figure 2). The actors and actresses in the video prompt were native speakers of English. During the recording, the researcher directed the speed of reading, tone of voice, and use of non-verbal actions (i.e., gestures, facial expressions, etc.) to prevent any unwanted variable affecting the difficulty of the items. The use of spoken discourse features (i.e., fillers, backchannels) was controlled to a minimum but use of gestures and facial expressions were allowed if they were natural. Both the prompt and the item were provided in a CBT format to prevent the split attention effect. Specifically, the video prompt was presented on the top half of the screen, and the question (along with the answer choices) were presented at the bottom half of the screen. This way, test-takers did not have to navigate to a separate window to view the question and the answer choices. For the audio-only group, the positioning of the prompt and the item features was equivalent to the video group but the space for the video was covered with a grey box showing a short phrase: 'listen to the passage'(see Figure 3).

Figure 3 Screenshot of the audio-only test

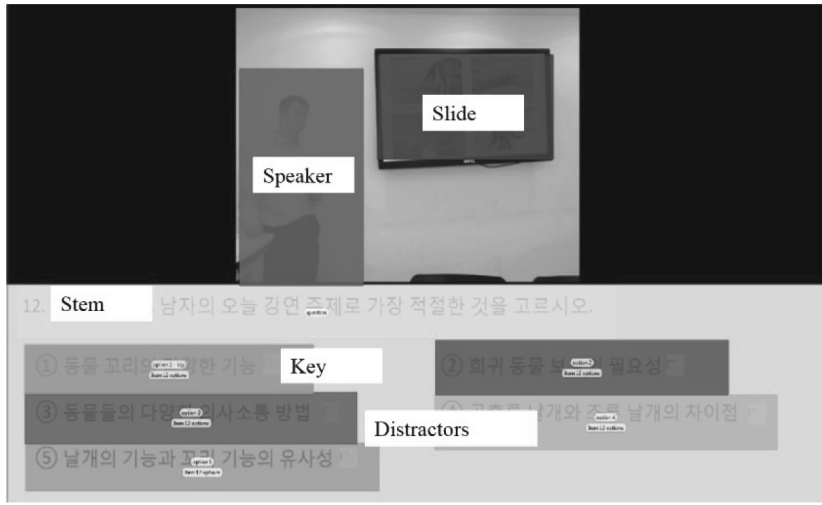


Data collection

Each test-taker sat the same listening test in either the video or audio-only mode individually. To measure test-takers' eye movements, a Tobii X2-60 eye tracker was used, which has a sampling rate at 60Hz, 45–90cm operating range between the eye-tracker and test-takers, and freedom of head movement at 70cm. The eye tracker was installed in the centre of the bottom frame of a 17-inch laptop to prevent any distractions to test-takers. All eye-tracking tests began with a nine-point calibration phase in order to ensure the quality of eye movements. Test-takers' interactions with the computer screen, such as their mouse movements and clicks, were recorded simultaneously with their eye movements in Tobii Studio. The time allowance given for each item was strictly controlled so the test would advance to the next item automatically. This helped prevent test-takers from accidentally advancing to the next item by either clicking or pressing the 'enter' key. An equivalent length of pauses between the instruction and the prompt, and between items, were provided as a default setting. The participants' test responses were manually coded as binary data (0: incorrect, 1: correct) and as aggregated test scores (maximum=15).

Test-takers' moment-to-moment eye movements on the Areas of Interest (AOIs) were measured and processed via Tobii Studio. The data in either seconds (s) or frequency (count) unit were exported from Tobii Studio for analysis. The AOIs are defined as specific areas that are assigned within the screen that a researcher is interested in (Holmqvist et al 2011). In total there were two visual AOIs: speaker and PowerPoint slides, which contain pictorial/textual resources that are semantically congruent with the talk – only applicable to the lecture-type items. There were three item AOIs: stem (task instructions), key (specified area for the correct 'key' option among five multiple-choice options of each item), and distractors (specified areas for the four incorrect 'distractor' options among five multiple-choice options of each item). All of these AOIs were activated from the beginning to the end of each item. The number of times and amount of time (both single fixations and total fixations) test-takers attended to these AOIs were measured and analysed to determine the degree to which they received information from these AOIs while completing the listening tasks (see Figure 4). It should be noted that the distractors AOI is a group of AOIs that consisted of four distractors. For this reason, the fixation duration measure is presented as the mean values of the four distractors of each item (i.e., fixation duration distractor = (option 1 + option 2 + option 3 + option 4)/4). This was essential as they were durations of each single fixation (fixation duration) on an AOI before moving on to another AOI.

Figure 4 Sample view of the AOIs



Data analysis

To answer RQ1, test-takers' overall test scores (maximum=15) and individual item scores (binary, 0 and 1) were compared between the two groups: video and audio-only. To examine the effect of video for each item separately, relative scores of each item (between 0 and 1) which indicate the proportion of items answered correctly were also compared between the two groups. To further examine whether video has a significant effect on test-takers' performance on each item, Mann-Whitney U tests were carried out on all 15 items separately.

For eye-movement analysis, test-takers' eye-movement data was calculated in three measures: 1) fixation count, 2) fixation duration, and 3) proportion of total fixation duration. The descriptions of these metrics are presented as follows: the first two metrics (count and duration) are raw eye-movement data extracted directly from the Tobii Studio, but the 'proportion of total fixation duration' is a relative value calculated from the raw data (see Table 1). The eye-tracking data also provides 'an indication of our referential decision', based on the assumption that a listener looks at what they have heard (Conklin et al 2018:114). Specifically, longer durations and frequent fixations on an item may indicate the listener is spending a large amount of effort to process the input (Pickering, Frisson, McElree and Traxler 2004).

To answer RQ2, descriptive statistics of the eye-movement data were presented first to illustrate test-takers' overall viewing patterns. To further

Table 1 Eye-movement measures

Metrics	Unit	Descriptions
Fixation count	Number of times	Number of times one fixates on an AOI. This is the sum of the number of times participants fixate on the AOI including the returning fixations to the same AOI.
Fixation duration	Seconds (s)	Average duration of all fixations on an AOI. Fixation duration indicates how long one's eye movement dwells on an AOI.
Proportion of total fixation duration	Percentage (%)	Proportion of time one fixates on an AOI. This is measured by calculating total fixation duration on an AOI divided by total fixation duration on all AOIs.

investigate the differences in viewing behaviours by test conditions, test-takers' eye-movement data were compared between the audio-only group and the video group for all three measures using either an independent samples t-test or a Mann-Whitney U test.

Results

The effect of videos on test-takers' overall test scores

The first RQ was about the extent to which adding a video prompt in a listening test has an effect on test-takers' listening test performance. The equivalence of the baseline NELT scores between the two groups was examined using a non-parametric Mann-Whitney U test as the score distribution violated the normality assumption. The two groups were found not significantly different in their listening ability ($U = 1645.5$, $p = 0.722$, $r = 0.03$).

Descriptive statistics of test-takers' listening scores under two different prompt conditions are presented in Table 2. It can be seen that test-takers in the video group scored higher than the test-takers in the audio-only group. A non-parametric Mann-Whitney U Test was conducted again to find that this difference was statistically significant ($U = 1307$, $p < 0.05$, $r = 0.20$).

Table 2 Descriptive statistics of final listening test scores

	Group	N	Mean	SD	K-S Test Sig[1]
Total score (Max = 15)	Audio	57	10.40	2.80	0.20
	Video	60	11.50	2.50	< 0.01

[1] *A Kolmogorov-Smirnov Test significance value smaller than .05 suggests a violation of the normality assumption.*

Descriptive statistics of item-level scores (0 = incorrect, 1 = correct) suggest that the scores of the video group were higher than those of the audio-only group in all items but items #1 and #4. However, the Mann-Whitney U tests on each item only found significant score differences in two items (#2 and #6, shaded in Table 3). That is, the video prompt only had a significant positive effect on score improvement in two out of the 15 items. It should be noted that item #2 is a dialogue type (inferencing task) and item #6 is a lecture type (finding specific information task).

In summary, test score analyses showed that there is a significant positive effect of visual input on L2 test-takers' overall listening performance. However, this effect is relatively weak and limited in that the significant differences were only found in two items.

Test-takers' eye movements under the two different conditions

RQ2 addressed the issue of the extent to which test-takers' viewing behaviours were associated with the presence of visual input. To answer this question, the eye-movement data on the three common AOIs – stem, distractors, and key – were compared between the two groups. The eye-movement data on the two visual AOIs – the speaker and the PowerPoint slides – were not compared as they were only applicable to the video group. Instead, descriptive statistics of the eye movements on these two visual AOIs were presented. Due to failure of the calibration, seven test-takers were excluded from the eye-tracking analysis.

As shown in the descriptive statistics presented in Table 4, the audio-only group fixated on all of the three AOIs (stem, key, and distractors) more frequently than the video group, but the mean single fixation duration on the AOIs was generally similar between the two groups.

For fixation count data, an independent samples t-test was conducted for pairwise comparisons on the three AOIs – stem, distractors, and key.

Table 3 Analysis of individual item

No.	Audio-only (N = 57)	Video (N = 60)	*U*	No.	Audio-only (N = 57)	Video (N = 60)	*U*
Item 1	0.21	0.15	1813.5	Item 9	0.79	0.81	1663.5
Item 2	0.54	0.73	1386*	Item 10	0.86	0.90	1641
Item 3	0.63	0.77	1479	Item 11	0.43	0.53	1518
Item 4	0.75	0.73	1746	Item 12	0.84	0.97	1525.5
Item 5	0.95	0.95	1705.5	Item 13	0.84	0.92	1582.5
Item 6	0.81	0.97	1437**	Item 14	0.79	0.90	1521
Item 7	0.72	0.85	1486.5	Item 15	0.53	0.62	1555.5
Item 8	0.68	0.72	1813.5				

*$p < .05$, ** $p < .01$

Table 4 Descriptive statistics of fixation count

	Group		N	Min	Max	Mean	SD	K-S Test Sig.[1]
Fixation count (unit: frequency)	Stem	Audio	53	7.33	91.87	51.04	20.29	0.20
		Video	57	15.80	54.27	29.21	8.30	0.20
	Distractors	Audio	53	13.40	156.80	87.54	33.95	0.20
		Video	57	15.40	83.00	47.70	14.79	0.20
	Key	Audio	53	3.60	52.33	25.09	11.49	0.20
		Video	57	2.73	30.40	14.58	5.68	0.20
	Speaker	Video	57	42.80	177.27	80.56	25.33	0.20
	slide	Video	57	13.00	90.43	42.30	17.71	0.20
Fixation duration (unit: seconds)	Stem	Audio	53	0.13	0.29	0.19	0.04	0.02
		Video	57	0.14	0.25	0.19	0.02	0.01
	Distractors	Audio	53	0.11	0.28	0.19	0.04	0.00
		Video	57	0.10	0.26	0.19	0.03	0.20
	Key	Audio	53	0.09	0.32	0.20	0.05	0.01
		Video	57	0.12	0.34	0.20	0.05	0.04
	Speaker	Video	57	0.22	1.16	0.54	0.24	< 0.01
	slide	Video	57	0.16	0.41	0.26	0.06	0.20

[1] *A Kolmogorov-Smirnov Test significance value smaller than .05 suggests a violation of the normality assumption.*

Findings suggest that the differences in the number of fixations between the two groups were all statistically significant. As presented in Table 5, the significant large gaps found in fixation count data between the two test conditions were expected since the video group was provided with additional visual input. Under the same total test duration, having the video prompts may have naturally made the video group look less frequently at the question and the options. In fact, test-takers in the video group viewed the speaker and the slide AOIs many times (see Table 4).

To compare the fixation duration data, median values were used instead because the data were not normally distributed. Findings of the Mann-Whitney U test suggest that the median fixation duration values for all three AOIs were not significantly different between the two groups (see Table 6). It should be noted that test-takers in the video group made relatively longer single fixations to the speakers (M = 0.5 seconds) and the slide (M = 0.3 seconds) than other AOIs (see Table 5).

Descriptive statistics of the proportion of total fixation duration are presented in Table 7 below. Findings suggest that the audio-only group spent a greater proportion of time viewing all three AOIs than the video group.

Findings of independent samples t-tests as shown in Table 8 suggest that the differences found in the proportion of total fixation duration on all three AOI variables were statistically significant. The large differences found between the two groups were expected because test-takers in the video group

Table 5 Independent samples t-test result on fixation count data

AOIs	t	*df*	Sig. (2-tailed)	Mean difference
Stem	7.29	67.90	< 0.001	21.83
Distractors	7.88	69.95	< 0.001	9.96
Key	6.01	74.72	< 0.001	10.51

Table 6 Mann-Whitney U test results on fixation duration

AOIs	*U*	*p-value*	Median difference	*r*
Stem	1488	0.89	–0.003	0.01
Distractors	1384	0.45	0.008	0.01
Key	1486	0.88	0.013	0.07

had both visual cues and stem and options to view while the audio-only group only had stem and options.

To investigate the proportion of time viewing the visual AOIs and compare it with the three item AOIs, descriptive statistics of the video group's proportion of total fixation duration data on all AOIs are presented below. In this analysis, the proportion of total fixation duration values were recalculated and presented for the dialogue prompts and the lecture prompts separately because the AOI 'PowerPoint slide' is only present in the lecture items. This was essential since the 'proportion' of the total fixation

Table 7 Differences in proportion of total fixation duration (Unit: %)

	AOIs	Group	N	Min	Max	Mean	SD	K-S Test Sig.[1]
Proportion of total fixation duration	Stem	Audio	53	9.93	62.62	32.66	11.25	0.16
		Video	57	1.11	26.50	10.73	4.35	0.09
	Distractors	Audio	53	30.80	67.35	52.28	8.41	0.20
		Video	57	3.65	35.41	18.63	6.72	0.20
	Key	Audio	53	6.57	25.14	15.06	4.18	0.20
		Video	57	0.92	12.82	5.79	2.61	0.20

[1] *A Kolmogorov-Smirnov Test significance value smaller than .05 suggests a violation of the normality assumption.*

Table 8 Independent samples t-test results on proportion of total fixation duration

AOIs	*t*	*df*	Sig. (2-tailed)	Mean difference
Stem	13.29	66.28	< 0.001	21.93
Distractors	23.26	108.00	< 0.001	33.65
Key	13.84	86.01	< 0.001	9.28

duration is interpreted on each AOI out of 100% and the two prompt types have different numbers of AOIs. Hence, the mean proportion values of each prompt type sum up to 100 (see Tables 9 and 10, and Figure 5).

Descriptive statistics show that test-takers in the video group gazed on average at the speaker(s) approximately 61% of the time in the dialogue items, and 50% of the time on the speaker(s) and 19% of the time on the PowerPoint slides in the lecture items. The standard deviation and the skewness and kurtosis values all suggest that the distribution of the proportion of total fixation duration values was generally normal. In addition, it should be noted that the standard deviation values of the visual AOIs are generally bigger than those of the item AOIs (see Tables 9 and 10). Therefore, it can be argued that the reliance on the video prompts varies across individual test-takers.

In summary, test-takers in the video group viewed the visual cues in the video more than half of the entire viewing duration in the test, which therefore caused the significant difference in the proportion of total fixation duration values between the two test conditions. To verify this finding, an additional analysis was carried out to investigate the differences in the proportion of total fixation duration on the three common AOIs only. In the analysis, the proportion of total fixation data of the video group was recalculated by replacing the denominator with the sum of total fixation duration on the three common AOIs only, which is identical to the denominator for the audio-only group. Through this analysis, it was possible to directly compare the proportion of fixation duration on the three common AOIs between two subgroups. The descriptive statistics of this analysis are presented in Table 11 below.

Table 9 Proportion of total fixation duration: dialogue

AOIs	N	Min	Max	Mean	SD	Skewness	Kurtosis
Stem	57	1.43	29.49	11.91	5.13	0.74	1.62
Distractors	57	4.60	41.42	21.00	7.66	0.15	0.26
Key	57	1.19	13.30	6.02	2.99	0.64	0.17
Speaker	**57**	**36.46**	**92.68**	**61.06**	**11.43**	**0.32**	**0.44**

Table 10 Proportion of total fixation duration: lecture

AOIs	N	Min	Max	Mean	SD	Skewness	Kurtosis
Stem	57	0.74	23.09	9.38	4.72	0.71	0.88
Distractors	57	2.02	34.73	15.92	7.16	0.40	0.27
Key	57	0.62	12.91	5.52	2.82	0.51	0.14
Speaker	**57**	**17.61**	**69.27**	**49.95**	**10.47**	**–0.37**	**0.75**
Slide	**57**	**6.58**	**40.69**	**19.23**	**7.67**	**0.76**	**0.52**

Figure 5 Pie charts for proportion of total fixation duration

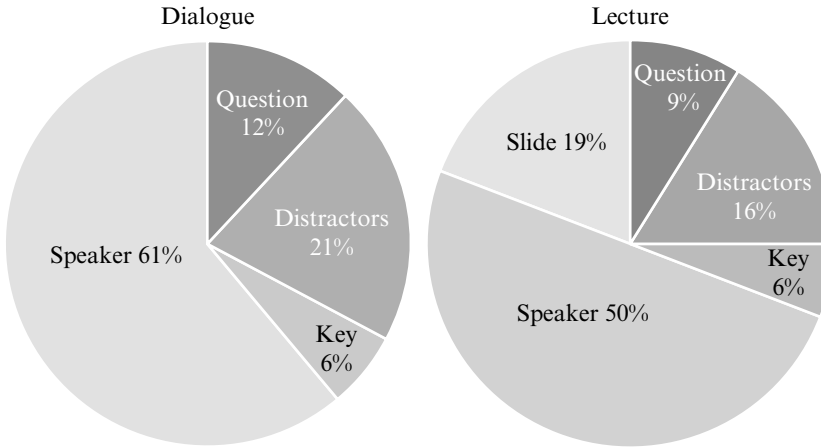


Table 11 Descriptive statistics of the proportion of total fixation duration: common AOIs only

Group		N	Mean	Median	SD	K-S Test Sig.[1]
Stem	Audio	53	32.66	30.60	11.25	0.16
	Video	57	31.94	31.11	10.45	0.20
Distractors	Audio	53	52.28	53.74	8.41	0.20
	Video	57	50.35	50.44	8.16	< 0.05
Key	Audio	53	15.06	14.85	4.18	0.20
	Video	57	16.78	16.51	5.73	< 0.05

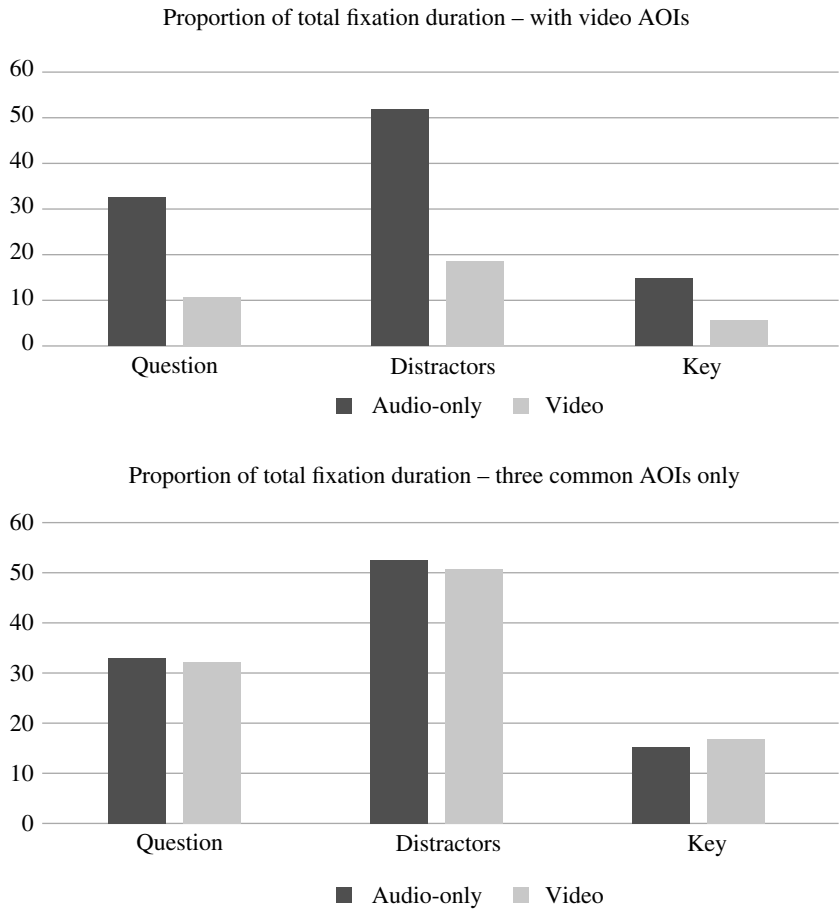
[1] *A Kolmogorov-Smirnov Test significance value smaller than .05 suggests a violation of the normality assumption.*

Unlike the results shown in Table 7, the differences found in the proportion of total fixation duration on the three AOIs are no longer substantial when the visual AOIs are removed from the calculation. Findings of the independent samples t-test suggest that there were no statistically significant differences in the proportion of total fixation duration on the three common AOIs between the two groups (see Table 12). Since the video group's data for the distractors and the key were not strictly normally distributed, Mann-Whitney U tests were carried out. The results were consistent with those from the independent samples t-test as no statistical differences between the two groups were found in viewing the distractors ($U = 1295$, n.s., $r = 0.12$) and the key ($U = 1231$, n.s., $r = 0.16$). These findings suggest that the proportion of viewing time on the stem, the key, and the distractors was not significantly

Table 12 Independent samples t-test for proportion of total fixation duration

	T	df	Sig.(2-tailed)	Mean dif	SE dif	95% CI	
						Lower	Upper
Stem	0.34	105.72	0.73	0.71	2.07	–3.40	4.83
Distractors	1.22	106.86	0.23	1.93	1.58	–1.21	5.06
Key	–1.80	102.34	0.08	–1.71	0.95	–3.60	0.18

Figure 6 Bar graphs of proportion of total fixation duration (unit: %)



different between the two subgroups when the data for visual AOIs were taken out (see Figure 6). This suggests that the large differences found in the proportion of total fixation duration on the stem, the distractors, and the key AOIs between the two test conditions were attributed to the presence of the visual cues.

Discussion

Regarding the effect of visual input on listening performance, this study has shown that the test-takers in the video group performed slightly better than those in the audio-only group. However, it should be noted that the mean score difference was not large (1.10 total out of 15) and the effect size of the significant difference was also small (r = .20) according to Plonsky and Oswald's (2014) standard. This finding was in line with a few previous studies which also reported a positive effect but small effect sizes (Wagner 2010, 2013, Wilberschied and Berman 2004). Therefore, it cannot be concluded that the video mode had a strong facilitating effect on L2 listening test performance. In addition, there may be an interaction effect between proficiency level and prompt mode on L2 listening test performance. In summary, having the visual input aids test-takers in comprehending the aural input better but this does not necessarily mean that a video-mediated listening test is substantially easier than an audio-only test. Whether employing the video test format might be a better choice or not needs further investigation considering the intended construct and test-taker characteristics.

The study also found a weak facilitating effect of visual input on test performance when the item features (stem and answer choices) are provided next to the video prompt. This test design is different from that adopted by previous studies in which stem and answer choices are presented separately (Batty 2015, Cubilo and Winke 2013, Suvorov 2015, Wagner 2010). As aforementioned, presenting both the video prompt and stem and answer choices together on the same screen was more effective in assessing test-takers' performance as it may prevent a split-attention effect or memory effect separately (Wagner 2010, 2013). The test design also allowed this study to measure the proportion of time spent on viewing the visual input and the item features. Such eye-movement analysis provides a better understanding of L2 learners' cognitive processes during test-taking.

It is also important to note that although the test used in this study had a lower number of items, it had more listening prompts (15) and listening scenarios compared to the previous studies (see Table 13). In addition, presenting only one item per each listening prompt was more effective because it required less cognitive effort. Other existing studies used longer

Table 13 List of existing studies: video mode outperforming audio-only mode

Work	No. of items	No. of prompts	N	Type of prompt	Effect size
Parry and Meredith (1984)	60	27	178	Dialogue	0.75, 0.99, 1.10 (by proficiency)
Shin (1998)	18	4	83	Lecture	$d = 0.93$
Sueyoshi and Hardison (2005)	4	1	42	Lecture	Not reported
Wagner (2010)	40	6	202	Dialogue, Lecture	$d = 0.35$
Wagner (2013)	26	4	192	Dialogue, Lecture	$\eta^2 = 0.04$
Wilberschied and Berman (2004)	14 (4~5 multiple-choice items each)	14	61	Monologue (broadcast), Cartoons	$R^2 = 0.30, 0.13$ (by grade)

listening prompts with more items under each prompt. It should be noted that listening to a longer prompt and answering multiple questions demands a higher cognitive load in the working memory (Buck 2001, Field 2015).

As presented above, the positive influence of visual input on listening test performance was found in many previous studies conducted in different contexts with different foci. Findings of these studies together showed that the presence of video makes the listening test slightly 'easier' compared to the traditional audio-only listening test. However, one can argue that a test being 'easier' does not necessarily mean it is a 'better' test. This is why further investigation on specific causes of improved listening test performance is required. We may interpret this phenomenon based on a connectionist view of cognitive processes involved in listening comprehension (Li 2013:64). The connectionist view suggests that the interpretations of aural and visual input are intertwined in one's cognitive system and help listeners to construct meaning using their background knowledge (Guichon and McLornan 2008, Lynch 2009). This study also supports the multimedia learning theories, namely Mayer's cognitive theory of multimedia learning (2009) and Paivio's dual coding theory (1990). Both theories argue that visual and aural input are processed separately in our cognitive system and dual presentation of these two types of input enhances comprehension and language learning.

The study found an impact of the presence of visual prompts on test-takers' viewing patterns. Specifically, it was found that the audio-only group generally viewed the stem and the answer choices longer and more frequently than the video group which received extra visual input. In fact, findings showed that test-takers in the video group attended to the visual input for approximately 61% (dialogue) and 69% (lecture) of the entire viewing time. This is consistent with Wagner's (2007) finding that test-takers view the

video for 69% of the time. While Wagner (2007) only reported the amount of time test-takers attended to the entire test screen, this study provided more detailed information about the amount of time fixated on each specific visual cue, with the help of the eye-tracking technology. Similarly, a previous eye-tracking study found that the proportion of time viewing the images was higher than that of reading subtitles (text) when watching a video (Mestres and Pellicer Sánchez 2019). It is evident that learners' viewing patterns are significantly affected by the mode of input, particularly in terms of how visual cues are presented to them.

Taken together, it can be argued that test-takers' listening processes could be significantly different between the video and the traditional audio-only conditions. Despite different viewing behaviours, the listening test scores between the two test conditions were not largely different. In other words, the presence of video attracted test-takers' attention to the visual input but this only had a weak positive effect on their listening test performance.

Furthermore, the observation that the audio-only group had longer and more frequent fixations on the stem and the answer choices may indicate that they devoted more effort to reading and deciding on an answer. Such cognitive effort in selecting an answer in a multiple-choice test can be seen as a test-wiseness strategy, where test-takers simply match what they hear with the textual information provided in the options to 'eliminate' wrong answers (Cohen 2007, Lee and Winke 2013, Teng 2013). Conversely, shorter and less frequent fixations on the stem and answer choices in the video group may suggest less use of such strategies when the video was given, presumably because they spent a considerable amount of time viewing the video prompt.

It should be also noted that there was no significant difference in the mean single fixation duration on the item AOIs between the two groups. In theory, longer fixation duration indicates a higher level of cognitive demand in comprehending text (Rayner 1998, Rayner and Pollatsek 2013). Based on this theory, it can be argued that test-takers in both groups spent a similar amount of cognitive effort in selecting an answer option. In addition, it was also found that the mean single fixation durations on the speaker and the PowerPoint slide were generally longer than the mean fixation durations on the stem and answer choices (see Table 4). Unlike the underlying assumption behind reading text, visual attention is believed to be driven by the 'saliency', 'informativeness', and 'unusualness' of the given input (Conklin et al 2018:115). It can be speculated that the test-takers' attention was attracted to the PowerPoint slides in the lecture items because there were novel animation features and possibly clues to an answer.

Findings of the test score comparisons and the eye-movement analysis together showed that the video group test-takers spent less time reading the stem and the answer choices but still performed slightly better than the

audio-only group. To some extent, this finding suggests that as long as the test-takers are given a reasonable amount of time to figure out what the task is about (reading the stem) and what options to choose from, viewing them for longer or more frequently may not necessarily improve the overall test performance dramatically. In addition, viewing the visual input for longer did not massively improve test-takers' listening test performance. Instead, the longer attention paid to the visual input only had a weak positive effect on the overall test score, and significant score differences were found in two out of 15 items only.

With this in mind, it may be argued that the main reason why the test-takers in the video group performed slightly better than those in the audio-only group is largely due to the additional information provided through the visual input. Although the additional visual information might not be decisive clues leading to an answer, it may have aided the test-takers in understanding the context of the speech better.

In terms of the cognitive processes of viewing the video prompt, descriptive statistics of eye-movement data on the speaker and the PowerPoint slide showed that test-takers in the video condition spent more than half of the time viewing the visual input in both the dialogue and the lecture items. On the one hand, it can be speculated that test-takers in the video group were actively processing the visual input because they found it useful. On the other hand, it might be simply because they had enough time to read the stem and the options and thus spent the remaining time viewing the video prompt. It should be noted that the standard deviations of the proportion of total fixation duration values on the visual AOIs were all bigger than those of the item AOIs. This finding suggests that the extent to which test-takers relied on the visual input generally varied by individual test-taker. This supports the concerns about the construct-irrelevant variance caused by the individual differences in utilising the visual input in the listening comprehension process (Buck 2001, Burgoon 1994, Wagner 2008). However, whether individual difference is a construct-relevant factor or not remains debatable because individual difference could also mean that the test-takers use various test-taking strategies. Suvorov (2018) argues that test-takers' individual differences in utilising visual input can be seen as a construct-relevant element if they reflect real-life listening scenarios. More studies that investigate the use of different strategies in a video listening test are required to examine closely how test-takers specifically use the visual input as part of their test-taking strategies.

Findings of this study have shed light on the possibility of introducing video prompts in L2 listening comprehension tests based on more specific and up-to-date listening test constructs. Similar positive effects of video can be found in other language tests that involve a listening section, particularly the ones that contain similar prompt types (monologue and dialogue), response types (multiple choice), and the target participants (EFL) to this

study. However, to propose a new construct specifically designed for a video-mediated listening test, further systematic inquiries on the ways in which the presence of visual aids have an impact on test-takers' listening comprehension process and performance are needed to support valid arguments. Also, test developers need to consider possible random effects caused by discrete characteristics of the item and/or the test-takers.

Conclusion

This study investigated the effect of visual input provided in a listening test on test-takers' listening performances and viewing behaviours. Based on the study findings, I argue that adding visual input to a traditional audio-only listening test is desirable.

Using participants' eye-movement data as evidence of cognitive processes is still a relatively new approach to validating L2 listening tests. The use of eye-tracking technology allows researchers to unveil test-takers' cognitive processes when exposed to multimodal input. In this regard, this study appears to be one of the few early studies of this method.

However, this study is limited in that the eye-movement data were not triangulated with other sources of qualitative data (e.g., a retrospective verbal report), as recommended by many existing eye-tracking studies (Bax and Chan 2019, Conklin et al 2018, Guan, Lee, Cuddihy and Ramey 2006). Questions remain as to why the test-takers showed different viewing behaviours when the video prompt was present. Combining eye tracking with stimulated-recall interviews (as in Suvorov 2018) is strongly recommended for future studies that investigate this topic. By doing so, the rationales and motives behind test-takers' viewing patterns can be interpreted more confidently (see Kwon and Yu 2024). In addition, the present study did not consider participants' language proficiency in the research design. Also, presenting the stem and answer choices in the test-takers' L1 may also have affected their test-taking processes and performances. This was a practical decision made to accommodate the participants with the test format they were familiar with.

References

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

Bachman, L F and Palmer, A S (2010) *Language Assessment in Practice: Developing Language Assessments and Justifying Their Use in the Real World*, Oxford: Oxford University Press.

Batty, A O (2015) A comparison of video- and audio-mediated listening tests with many-facet Rasch modeling and differential distractor functioning, *Language Testing* 32 (1), 3–20.

Batty, A O (2017) *The impact of visual cues on item response in video-mediated tests of foreign language listening comprehension*, unpublished doctoral dissertation, Lancaster University, available online: eprints.lancs.ac.uk/id/eprint/84467/4/batty_thesis_news.pdf

Batty, A O (2018) Investigating the impact of nonverbal communication cues on listening item types, in Ockey, G J and Wagner, E (Eds) *Assessing L2 Listening: Moving Towards Authenticity*, Amsterdam: John Benjamins Publishing Company, 161–175.

Batty, A O (2020) An eye-tracking study of attention to visual cues in L2 listening tests, *Language Testing* 38 (4), 511–535.

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

Bejar, I, Douglas, D, Jamieson, J, Nissan, S and Turner, J (2000) *TOEFL 2000 Listening Framework: A Working Paper*, available online: www.ets.org/Media/Research/pdf/RM-00-07.pdf

Buck, G (2001) *Assessing Listening*, Cambridge: Cambridge University Press.

Burgoon, J K (1994) Nonverbal signals, in Knapp, M L and Miller, G R (Eds) *Handbook of Interpersonal Communication Volume 2*, Thousand Oaks: Sage Publications, 229–285.

Cohen, A (2007) The coming of age for research on test-taking strategies, in Fox, J, Wesche, M and Bayliss, D (Eds) *Language Testing Reconsidered*, Ottawa: University of Ottawa Press, 89–111.

Coniam, D (2001) The use of audio or video comprehension as an assessment instrument in the certification of English language teachers: A case study, *System* 29 (1), 1–14.

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

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

Cubilo, J and Winke, P M (2013) Redefining the L2 listening construct within an integrated writing task: Considering the impacts of visual-cue interpretation and note-taking, *Language Assessment Quarterly* 10 (4), 371–397.

Elliott, M and Wilson, J (2013) Context validity, in Geranpayeh, A and Taylor, L (Eds) *Examining Listening: Research and Practice in Assessing Second Language Listening*, Studies in Language Testing Volume 35, Cambridge: UCLES/Cambridge University Press, 152–241.

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

Field, J (2015) *The effects of single and double play upon listening test outcomes and cognitive processing*, available online: www.britishcouncil.org/exam/aptis/research/publications/effects-single-and-double-play

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

Ginther, A (2002) Context and content visuals and performance on listening comprehension stimuli, *Language Testing* 19 (2), 133–167.

Gruba, P (1993) A comparison study of audio and video in language testing, *JALT Journal* 15 (1), 85–88.

Guan, Z, Lee, S, Cuddihy, E and Ramey, J (2006) *The validity of the stimulated retrospective think-aloud method as measured by eye tracking*, paper presented at the SIGCHI Conference on Human Factors in Computing Systems, Montréal, April 22–27.

Guichon, N and McLornan, S (2008) The effects of multimodality on L2 learners: Implications for CALL resource design, *System* 36 (1), 85–93.

Hernandez, S (2005) *The effects of video and captioned text and the influence of verbal and spatial abilities on second language listening comprehension in a multimedia learning environment*, unpublished doctoral dissertation, New York University.

Holmqvist, K, Nyström, M, Andersson, R, Dewhurst, R, Jarodzka, H and Van de Weijer, J (2011) *Eye Tracking: A Comprehensive Guide to Methods and Measures*, Oxford: Oxford University Press.

Kwon, S K and Yu, G (2024) The effect of viewing visual cues in a listening comprehension test on second language learners' test-taking process and performance: An eye-tracking study, *Language Testing*. doi. org/10.1177/02655322241239356

Lee, H and Winke, P M (2013) The differences among three-, four-, and five-option-item formats in the context of a high-stakes English-language listening test, *Language Testing* 30 (1), 99–123.

Lesnov, R O (2018) *The role of content-rich visuals in the L2 academic listening assessment construct*, PhD thesis, Northern Arizona University.

Li, Z (2013) The issues of construct definition and assessment authenticity in video-based listening comprehension tests: Using an argument-based validation approach, *International Journal of Language Studies* 7 (2), 61–82.

Londe, Z C (2009) The effects of video media in English as a second language listening comprehension tests, *Issues in Applied Linguistics* 17 (1), 41–50.

Lynch, T (2009) *Teaching Second Language Listening*, Oxford: Oxford University Press.

Mayer, R E (2009) *Multimedia Learning* (Second edition), New York: Cambridge University Press.

Mestres, E T and Pellicer Sánchez, A (2019) Young EFL learners' processing of multimodal input: Examining learners' eye movements, *System* 80, 212–223.

Muñoz, C (2017) The role of age and proficiency in subtitle reading. An eye-tracking study, *System* 67, 77–86.

Ockey, G J (2007) Construct implications of including still image or video in computer-based listening tests, *Language Testing* 24 (4), 517–537.

Paivio, A (1990) *Mental Representations: A Dual Coding Approach*, New York: Oxford University Press.

Parry, T S and Meredith, R A (1984) Videotape vs. audiotape for listening comprehension tests: An experiment, *OMLTA Journal* 47–53.

Pickering, M J, Frisson, S, McElree, B and Traxler, J (2004) Eye movements and semantic composition, in Carreriras, M and Clifton, C (Eds) *Online Study of Sentence Comprehension: Eyetracking, ERP, and Beyond*, New York: Psychology Press, 33–50.

Plonsky, L and Oswald, F L (2014) How big is "big"? Interpreting effect sizes in L2 research, *Language Learning* 64 (4), 878–912.

Rayner, K (1998) Eye movements in reading and information processing: 20 years of research, *Psychological Bulletin* 124 (3), 372–422.

Rayner, K and Pollatsek, A (2013) *Psychology of Reading*, New York: Routledge.

Shin, D (1998) Using videotaped lectures for testing academic listening proficiency, *International Journal of Listening* 12 (1), 57–80.
Sueyoshi, A and Hardison, D M (2005) The role of gestures and facial cues in second language listening comprehension, *Language Learning* 55 (4), 661–699.
Suvorov, R (2009) Context visuals in L2 listening tests: The effects of photographs and video vs. audio-only format, in Chapelle, C A, Jun, H G and Katz, I (Eds) *Developing and Evaluating Language Learning Materials*, Ames: Iowa State University, 53–68.
Suvorov, R (2013) *Interacting with visuals in L2 listening tests: An eye-tracking study*, PhD thesis, Iowa State University.
Suvorov, R (2015) The use of eye tracking in research on video-based second language (L2) listening assessment: A comparison of context videos and content videos, *Language Testing* 32 (4), 463–483.
Suvorov, R (2018) Test-takers' use of visual information in an L2 video-mediated listening test, in Ockey, G J and Wagner, E (Eds) *Assessing L2 Listening. Moving Towards Authenticity*, Amsterdam: John Benjamins Publishing Company, 146–160.
Taylor, L (2013) Introduction, in Geranpayeh, A and Taylor, L (Eds) *Examining Listening: Research and Practice in Assessing Second Language Listening*, Studies in Language Testing Volume 35, Cambridge: UCLES/Cambridge University Press, 1–35.
Teng, H-C (2013) *Analysis of EFL learner' task strategies for a listening comprehension test*, paper presented at the KOTESOL, Seoul.
Vandergrift, L and Goh, C C (2012) *Teaching and Learning Second Language Listening: Metacognition in Action*, New York: Routledge.
Wagner, E (2007) Are they watching? Test-taker viewing behavior during an L2 video listening test, *Language Learning & Technology* 11 (1), 67–86.
Wagner, E (2008) Video listening tests: What are they measuring?, *Language Assessment Quarterly* 5 (3), 218–243.
Wagner, E (2010) The effect of the use of video texts on ESL listening test-taker performance, *Language Testing* 27 (4), 493–513.
Wagner, E (2013) An investigation of how the channel of input and access to test questions affect L2 listening test performance, *Language Assessment Quarterly* 10 (2), 178–195.
Weir, C J (2005) *Language Testing and Validation: An Evidence-based Approach*, Basingstoke: Palgrave Macmillan.
Wilberschied, L and Berman, P M (2004) Effect of using photos from authentic video as advance organizers on listening comprehension in an FLES Chinese class, *Foreign Language Annals* 37 (4), 534–540.
Winke, P M and Lim, H (2014) The effects of testwiseness and test-taking anxiety on L2 listening test performance: A visual (eye-tracking) and attentional investigation, *IELTS Research Reports Online Series* 3, 1–30, available online: www.ielts.org/~/media/research-reports/ielts_online_rr_2014-3.ashx
Winke, P M, Gass, S and Sydorenko, T (2013) Factors influencing the use of captions by foreign language learners: An eye-tracking study, *The Modern Language Journal* 97 (1), 254–275.

5 Investigating the cognitive validity of a reading test using eye-tracking technology and stimulated recall interviews

Nathaniel Owen
Open University, UK

Abstract

This chapter reports on a study which investigates whether test-takers use different item-completion processes for question types which assess different reading purposes. The research employs eye-tracking technology and stimulated recall interviews. Fourteen L2 English speakers at approximately Level C1 of the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001) completed an authentic Internet-based Test of English as a Foreign Language (TOEFL iBT®) Reading paper composed of basic comprehension and inferencing item types. Each participant then took part in a stimulated recall interview in which they were asked to elaborate on their item-completion processes. Eye-tracking and interview data revealed broadly similar item-completion processes, with participants using expeditious word spot strategies followed by careful local reading. However, inferencing question types elicited backtracking (backward sweeps) among participants more than basic comprehension items, indicating that the different reading purpose of this item type stimulated localised re-reading. The implications of the findings and recommendations for future eye-tracking studies are discussed with specific relevance to cognitive approaches to validation of reading tests.

Introduction

This chapter investigates how test-takers complete individual reading questions in a high-stakes test of reading comprehension using a combination of eye-tracking technology and stimulated recall interviews. Eye tracking is an increasingly popular methodology to investigate interactions between students and educational materials in a variety of academic disciplines. Examples include investigations of how students read materials (Rayner 1978, 1998), website functionality (Poole and Ball 2005), how and when

students refer to subtitles in video materials (Winke, Gass and Sydorenko 2013), interactions between students in online forums (Stickler, Smith and Shi 2016), and foreign language learning (Stickler and Shi 2017). This growth is also reflected in language testing with a new trend of studies employing this technology (e.g., Ballard 2017, Bax 2013, Suvorov 2015) in particular in relation to reading (e.g., Bax and Weir 2012, Brunfaut 2016, Brunfaut and McCray 2015). Despite the increasing number of studies investigating reading assessment using eye tracking, there has so far been little reflection on how data produced from this complex methodology contributes towards a wider cognitive validity research agenda. In most studies investigating reading, researchers examine the difference in eye-gaze patterns between successful and unsuccessful test-takers and claim that emergent differences show test tasks elicit responses consistent with expectations, therefore contributing towards a cognitive validity argument.

This chapter explores an alternative approach to investigating cognitive validity in tests of reading comprehension. Whereas previous studies (e.g., Bax 2013, Brunfaut and McCray 2015) have demonstrated that language proficiency of test-takers influences eye-gaze patterns, this chapter investigates whether test tasks which target different reading purposes can or should elicit different eye-gaze patterns independently of test-taker proficiency in the high-stakes Internet-based Test of English as a Foreign Language (TOEFL iBT). Based on the findings, the chapter offers some specific recommendations about the future of eye-tracking methods to investigate the validity of language tests, considering the drawbacks of the technology, in particular how the complexity of the methodology has led to difficulties in replication and building an established body of literature.

Literature

Cognitive validity of reading tests

A crucial component of modern validity theory proposed by Messick (1989) is cognitive validity, meaning that the mental processes elicited by test tasks should be consistent with our theoretical understanding of the mental processes required to perform successfully in the domain of interest (Weir 2005). For example, the kinds of reading required for successful completion of a test of reading for academic purposes should reflect the kinds of reading required for successful participation in higher education. This emphasis on cognitive validity has become integral to the field through inclusion in standards documents such as those for educational and psychological testing (e.g., American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME] and Joint Committee on Standards for Educational and Psychological Testing 2014). In order to

measure a psychological construct such as reading for academic purposes, it must be explicitly defined with careful consideration given to how it can be sampled and operationalised. The construct of reading for academic purposes may be associated with a variety of reading behaviours such as skimming, scanning, local and global reading and synthesising information across texts (Khalifa and Weir 2009, Weir 2005). However, individual test tasks cannot elicit behaviour which is fully representative of the construct due to logistical constraints. Therefore, validity claims for reading tests are usually made on the basis that different test tasks sample and operationalise different parts of the theorised construct. Buck (1991) argues that different types of test tasks therefore require different types of validity evidence.

Different test tasks are designed to elicit different cognitive elements of the construct. Koda (2005) and Khalifa and Weir (2009) outline cognitive models of reading which incorporate knowledge of vocabulary, grammatical systems and the ability to comprehend the ideas encoded into words, clauses, sentences, paragraphs and texts. Koda specifically identifies *text decoding*, *text-information building* and *situation-model construction*, characterised as three competencies: visual information extraction, incremental information integration, and text meaning, moderated by prior knowledge (Koda 2005:5). Tasks targeting text information-building should therefore present some form of divergent validity evidence that they do not assess text decoding (and vice versa). Khalifa and Weir's (2009) model includes a cognitive processing 'core'; a hierarchical model which begins at the orthographic, phonological and morphological word levels and progresses to higher-level mental representations of text content, including inferencing and building a mental representation of a text. As a result, Khalifa and Weir's model of reading suggests that tasks which assess inferential reasoning should elicit different cognitive processes than tasks which assess basic comprehension.

Reading purpose in the TOEFL iBT Reading test

The research reported in this chapter uses the TOEFL iBT Reading test because this test specifically identifies three reading purposes and labels each task accordingly: 'basic comprehension', 'inferencing' and 'reading-to-learn'. Basic comprehension describes the ability to comprehend ideas which are explicitly stated in the text. Inferencing tasks require test-takers to form connections between propositions which are not explicitly stated. Reading-to-learn tasks require test-takers to form connections between ideas across a text (Pearlman 2008:242–244). The test developers initially hypothesised that increasingly complex reading purposes would influence the cognitive demands placed upon test-takers, allowing test users to discriminate between stronger and weaker readers (Jamieson, Eignor, Grabe and Kunnan 2008:71). However, in the final test design, reading purpose is not

Table 1 Blueprint for TOEFL iBT Reading test with reading purposes and associated task types (Pearlman 2008:244)

Reading claim	**Test-taker can understand English language texts in an academic environment.**		
Reading sub-claim	**Basic comprehension** Can understand the lexical, syntactic and semantic content of the text and major ideas; can understand important sentence-level information; can connect information locally.	**Inferencing** Can comprehend an argument or idea that is strongly implied but not explicitly stated in the text; can identify the nature of the link between specific features of exposition and the author's rhetorical purpose; can understand the lexical, grammatical and logical links between successive sentences in a passage.	**Reading-to-learn** Can connect information across the entire text; can recognise the organisation and purpose of a text, understand the relative importance/ scope of ideas in a text; can understand rhetorical functions and purposes and organise (categorise/classify) important information into an appropriate mental framework representative of the organisation and inter-relationship of ideas in a text.
Task types	1. Vocabulary 2. Factual information 3. Sentence simplification 4. Pronoun reference	1. Inference 2. Rhetorical purpose 3. Insert text	1. Prose summary 2. Schematic table

intrinsically connected with task difficulty, as it is possible to develop easy reading-to-learn items and difficult tasks requiring test-takers to find discrete information. Different reading purposes are instead connected to different task types. This is outlined in the TOEFL iBT Reading blueprint which is reproduced in Table 1.

Cohen and Upton (2006) provide the most comprehensive attempt to capture data relating to cognitive processing in the TOEFL iBT Reading test. The authors collected verbal report data from 32 students who were assigned to complete six reading tasks. The authors used a combination of verbal protocol analysis and coded video-taped evidence of test-taker behaviour. Crucially, the authors concluded that test-takers did not use substantially different strategies to answer question types targeting different reading purposes, nor that reading-to-learn question types were more difficult than basic comprehension or inferencing question types. However, the operational definition of inferencing in the study is predicated on identifying the meaning of unknown lexis in context (2006:35). This definition does not capture the definition of inferencing in Table 1, which focuses on ideas and following internal logic of arguments across sentences.

Further investigation of the cognitive processing in the TOEFL iBT Reading test was undertaken by Owen (2016), who also used protocol analysis

with six test-takers using a coding schema developed from Khalifa and Weir's (2009) cognitive processing model of reading. This approach specifically sought to identify instances of inferential reasoning across propositions. This research provides evidence of inferential reasoning across clauses or sentences in 'inferencing', 'rhetorical purpose' and 'insert text' items. 'Inferencing' item types appear to prioritise a form of inferential reasoning known as 'bridging inferencing' (Singer 2007:346), in which the question stem provides a proposition and each of the options represents a possible conclusion of which only one may be logically derived using additional propositional information in the text. Many of the options in this item type contain anaphoric references back to the question stem providing further grounds for suggesting that 'anaphoric bridging inferencing' is the main understanding of 'inferential reasoning' encoded in TOEFL iBT inferencing items.

Further evidence from Owen (2016) demonstrates that multiple-choice tasks in the TOEFL iBT Reading test elicit a variety of cognitive processes. Inferencing and forming mental models of the text are associated with tasks that concentrate in the second half of the test. Basic comprehension tasks are based on establishing propositional meaning in single clauses or sentences and do not require inferencing or forming mental models of the text in order to complete them successfully, although Owen (2016) demonstrated that test-takers will sometimes 'over process' necessary task requirements to be certain of the answer. Owen's study also showed the weakness of relying only on protocol analysis as significant differences emerged between the findings and those of Cohen and Upton (2006). For example, Owen (2016) reported that eliminating incorrect options was a time-consuming strategy that test-takers only used whenever necessary to complete inferencing and reading-to-learn tasks, whereas the equivalent strategy in Cohen and Upton's study was reported as being used very frequently across task types.

Research question

This chapter now reports an empirical study designed to investigate task completion processes in the TOEFL iBT Reading test using eye tracking and stimulated recall interviews. The purpose is to investigate whether test-takers use different processes in different tasks depending on reading purpose. Currently, there is limited evidence in the literature that items designed according to different reading purposes elicit behaviour which is observably different using eye-tracking data. If differences can be observed between tasks which target different reading purposes, this data would strengthen a validity claim that the TOEFL iBT Reading test measures different reading purposes. Therefore, the research question (RQ) driving the study is: *What item-completion procedures do test-takers use to complete different task types in the TOEFL iBT Reading test?*

Methodology

Materials

A single authentic TOEFL iBT text and 13 associated items taken from official TOEFL test preparation materials (Educational Testing Service 2009) were selected on the basis that this research instrument contained items which had been previously shown to elicit a range of relevant cognitive processes (Owen 2016). Of the 13 questions, three were vocabulary items, five were factual information (including one negative factual information question), and there was one example each of sentence simplification, inference, rhetorical purpose, insert text and reading-to-learn (prose summary) items. As the research materials only included a single reading-to-learn task, this did not form part of the study as it would generate insufficient data to compare to tasks centred on other reading purposes. There was no example of a pronoun reference question associated with this text. The outline of this test is provided in Table 2 along with the claimed reading purpose for each task (see Table 1). Nine items assess basic comprehension and three items assess inferencing.

Eye-tracking instruments

The RQ was addressed using eye-tracking and stimulated recall interview data. Modern eye tracking is the precise measurement of eye activity in relation to an external stimulus, either screen-based or 'real-world' (Rayner 1998, Yarbus 1967). Infra-red cameras collect raw eye-movement data points up to 30–300 times per second, depending on the sampling rate of the eye tracker (for more information on various eye-tracking devices available on the market, see Conklin, Pellicer–Sánchez and Carrol 2018, Holmqvist et al 2011). The eye tracker used was a Tobii Pro X3-120 mobile device, which was attached to the bottom of a screen on a laptop computer that allowed participants to read naturally. The computer had a 19-inch monitor with a

Table 2 Test task characteristics

Text information	Item numbers	Question type	Reading purpose
660 words	1, 6, 8	Vocabulary	
	2, 4, 5, 7, 10	Factual information	Basic comprehension
	11	Sentence simplification	
Six paragraphs	9	Inference	
	3	Rhetorical purpose	Inferencing
	12	Insert sentence	
	13*	Prose summary*	Reading to learn*

**Item 13 did not form part of this study.*

resolution of 1,920 × 1,080 pixels. Eye-movement data was collected using the velocity threshold identification (I-VT) fixation filter (Komogortsev, Gobert and Jayarathna 2010), which relies upon velocity of eye movement to classify fixations, measured in visual degrees per second. The sample rate is 120 Hz per second, with the maximum angle between fixations set at 0.5 degrees, the maximum time between fixations at 75ms, and a minimum fixation duration of 60ms established to eliminate noise such as tremors or micro-saccades (Yarbus 1967) and to merge fixations that have been incorrectly split into multiple, shorter fixations which cannot be interpreted as processing of individual words.

Participants

To date, eye-tracking studies exploring second language reading tests have compared the behaviour of test-takers at different levels of proficiency. Bax and Weir (2012), Bax (2013), Brunfaut and McCray (2015), McCray and Brunfaut (2018) and Brunfaut (2016) have all demonstrated that eye-tracking data is sensitive to differences in test-takers' English language proficiency by comparing the differences in observable reading behaviour of successful and unsuccessful test-takers. However, the purpose of this research is to examine whether reading purpose might influence test-taker behaviour. As a result, participants needed to be as similar as possible to minimise the impact of proficiency on the data. Additionally, participants needed to be highly proficient to maximise the probability of recording correct responses, as only behaviour associated with correct responses can be used to make claims about the kinds of processing required for the different tasks.

Fourteen L2 English speakers were recruited for this study who were either postgraduate students or had undertaken university studies while based in the UK. All had been in the UK for at least one academic year and had achieved IELTS scores of 6.0 in speaking and reading prior to admission. Participants were familiar with English language tests and regularly used computers in their education or professional life. Eleven of the participants were Chinese, with one Russian, one French, and one German participant. The German participant and one Chinese participant were male, all other participants were female. Participants were also paid £20 in Amazon vouchers for participation.

Data collection and analysis

The test was uploaded to Tobii Studio Professional as a non-interactive pdf document. The layout of the document was kept as consistent as possible with how the test was presented in Educational Testing Service (ETS) documentation in order to maintain ecological validity. This ensured that

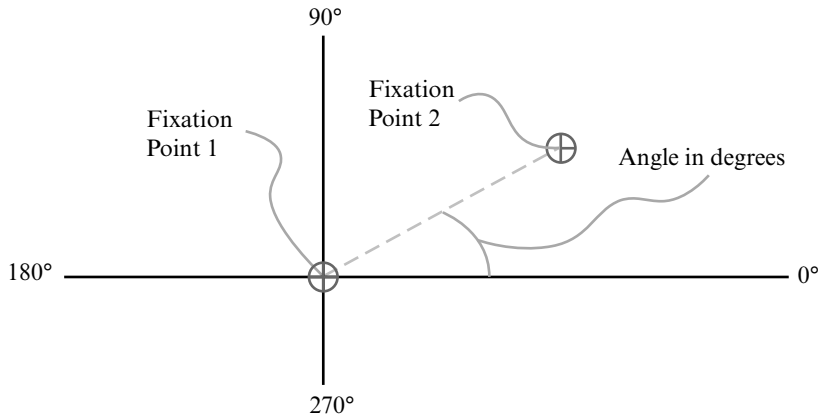
the number of words per line was consistent with the original sources. The font was Arial, size 14 point. As TOEFL iBT items mostly relate to individual paragraphs, the relevant paragraphs for each item were presented to participants alongside that item, consistent with source material design. Each question and paragraph were presented on individual pages. The task and text on each page were identified as separate areas of interest (AOIs) to separate text from question engagement. This was possible for all items except Item 12 (insert sentence). This item reproduces the paragraph four times with the relevant sentence shown in each of the possible four locations. For this reason, separate AOIs were created for each of the four paragraphs.

Participants were introduced to the test layout, and the procedure was explained to them verbally. All participants were presented with the eye-tracking technology prior to participation and were shown the types of data which emerge from engagement with onscreen stimuli. They were able to ask any clarifying questions about the procedure and the equipment used in the study and withdraw consent at any time. They were also reminded that their seated position as well as the position of the computer could not be moved once calibration had been undertaken. This is essential to avoid data loss in eye-tracking research in which participants are able to move freely. Each participant calibrated their eye movements using the in-built nine-point calibration tool in the Tobii Studio software with a distance from the screen of 50 centimetres. Participants then completed the reading test, marking their responses on a provided answer sheet. They were informed that the test would take about 20 minutes to complete, but the sessions were not timed, as it was deemed more important to the research aims to collect data for each item rather than have missing data due to time constraints. The researcher was present throughout the test, in the event of any technical difficulties. Eye-tracking data was stored locally in Tobii software to be downloaded later.

Methods of data analysis

Eye-tracking data was downloaded in raw format into an Excel file for each participant. Raw data is provided in the form of fixation duration (in milliseconds) and location (x and y co-ordinates for both eyes). Raw position signals from the left and right eye are combined into an average position signal to provide fixation co-ordinates. The location of each fixation is plotted on screen by counting the number of pixels vertically and horizontally from the bottom left-hand corner of the screen. Binary data (0,1) indicates which AOI each fixation is located in, allowing for aggregation of data relating to text or question. Raw data also includes absolute saccadic direction. This measures in degrees the location of one fixation relative to the fixation immediately prior (see Figure 1).

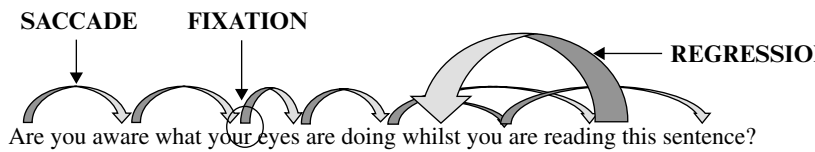
Figure 1 Absolute saccadic direction (Tobii Studio 2016:134)



We can then use basic geometry in Excel to calculate each saccade length in pixels, using the difference in x and y co-ordinates between two fixations and the absolute saccadic direction, measured in degrees. Saccades can therefore be forwards or backwards along a line of text. Backwards movements are also called backward sweeps. Regressive movements occur when the eyes backtrack along a line of text the reader has already read, in order to clarify cognitively difficult text, as depicted in Figure 2. Regressions, backwards motions for a distance of a few letters to reprocess an individual word, are examples of correcting 'inefficient' text processing (Rayner 1998). Backwards movements of fewer than 20 pixels were therefore eliminated from the research in order to focus on backward sweeps which return to different processing locations as these will be more indicative of processing difficulty.

Koda (2005:30) notes that eye-movement studies in L1 studies have demonstrated that each content word receives direct visual fixation, with function words omitted as the eyes jump to the next content word. Backward sweeps are identified when the difference in the x co-ordinate between two fixations is negative, indicating eye movement from right to left on the screen. Regression length can be calculated based on the same

Figure 2 Saccades, fixations and backward sweeps (Brunfaut and McCray 2015:10)



principles. Line returns are differentiated from backward sweeps by pixel length. This will vary from study to study depending on text size and screen resolution, but for this study, visual inspection of the data indicated that 400 pixels could be used as the initial cut-off to determine whether a backwards movement was a line return or a regression. The number of fixations, saccades and backward sweeps was weighted by the number of syllables in the target paragraph and question to account for differing text lengths for each question.

In order to compare eye-tracking data between 'basic comprehension' and 'inferencing' question types, six items, 2, 3, 6, 9, 11 and 12, were selected for closer analysis. Equal numbers of basic comprehension and inferencing questions were selected. Three items (2, 6 and 11) assessed basic comprehension (vocabulary, factual information and sentence simplification) and three (3, 9 and 12) assessed inferencing (inference, rhetorical purpose and insert text; see Table 2). All data for the analysis is selected from participants who responded correctly to these items. Test-takers who respond correctly are providing 'construct-relevant' text processing. That is, they are processing the test in a manner desired by the test developers. In contrast, test-takers who respond incorrectly are doing so because they are responding to the task in construct-irrelevant ways.

Additionally, ensuring the quality of the eye-tracking data recording is essential before data analysis can proceed. Sibley, Foroughi, Olson, Moclaire and Coyne (2017) note that data loss in eye-tracking experiments can vary dramatically, from 20% to 60% depending on the type of set-up used. Factors also associated with data loss include glasses and contact lenses (Holmqvist et al 2011) and pupil size (Drewes, Zhu, Hu and Hu 2014). The research design in this study requires participants to mark their responses on an answer sheet, resulting in the loss of tracking data when participants look down and away from the screen. Therefore, a pre-established cut-off point of capturing 70% of eye-gaze samples for each participant had to be met in order to include the data for analysis. The percentage is calculated as:

$$\text{Data loss} = 100 * \frac{N_{expected\,samples} - N_{valid\,samples}}{N_{expected\,samples}}$$

This figure is provided by Tobii as part of the data output (Tobii Studio 2016:92). 50% means that one eye or both eyes were found for half of the recording duration. In order to analyse the data, three specific hypotheses related to the RQ and the eye-tracking metrics were developed to compare items according to reading purpose. These are displayed in Table 3.

Table 3 Hypothesis statements

Metric	Hypothesis
Saccades (total number)	There is a statistically significant difference on the total number of saccades by reading purpose. As cognitive load increases so the overall number of saccades increases.
Fixation (total number)	There is a statistically significant difference on the total number of fixations by reading purpose. As cognitive load increases so the overall number of fixations increases.
Backward sweeps (total number)	There is a statistically significant difference on the total number of backward sweeps by reading purpose. As cognitive load increases, so the overall number of backward sweeps will increase.

Stimulated recall interviews

Immediately upon completion of the test, each participant undertook a stimulated recall interview (SRI), lasting approximately 20–30 minutes. SRI is a specific type of interview in which participants verbalise their thought processes after they have completed a task (Bowles 2010, Gass and Mackey 2000), stimulated by some element of task performance, and is a well-established research method for investigating L2 reading (Anderson 1991, Anderson, Bachman, Perkins and Cohen 1991, Bereiter and Bird 1985, Block 1986). Both Bax (2013) and Brunfaut and McCray (2015) employed this methodology alongside eye-tracking methodology. Eye-gaze patterns were overlaid on a video recording of the text and questions used in the test and were replayed to participants. This visual information acted as a stimulus for follow-up investigations in which the researcher asks the participant to provide verbal context to the eye-tracking data. The participants were able to pause the video and offer unsolicited thoughts at any moment. Thus, the interview was a collaborative effort by the participant and interviewer to co-construct the thought processes of the participant. If the participant remained silent, the interviewer offered prompts in relation to specific moments in the video. The interview protocol is outlined in Table 4.

Note that the statements and questions by the interviewer relate to specific concrete moments of the eye-gaze recording. These moments reflect metacognitive decision-making on the part of the test-taker. That is, they are conscious strategic decisions by the test-taker to either scan a text, read in more detail a specific part they have identified, or movements between the text and question to consider the options or meaning of the question. Focusing on these moments maximises the opportunity to obtain verbal data from the participants that explains their conscious decision-making. Additionally, it is important to focus on moments when test-takers select an option for a question. Once they have been 'primed' by reflecting on their

Table 4 Stimulated recall interview protocol

Observed behaviour of test-taker (eye-gaze data)	Verbal prompt by interviewer	Question asked by interviewer
Participant focuses on part of the text/question stem/option	'*I can see that you're focusing on* [insert part of question/text/task].'	'*Can you remember what you were thinking at that moment?*'
Participant moves quickly over part of the text/question stem/options	'*I can see that you're moving quickly over this part of the* [insert part of question/text/task].'	'*Can you remember what you were thinking at that moment?*'
Participant rereads part of a text (visual evidence of regression)	'*I can see that you reread this part of the* [insert part of text/task].'	'*Can you remember why you reread this part?*'
Participant hovers over one or more options in a question	'*I can see that you're considering* [this/these option/s].'	'*Can you remember what you were thinking about at this moment?*'
Participant moves quickly between the text and question one or more times	'*I can see that you're moving between these parts of the* [insert parts of question/text/task].'	'*Can you remember what you were thinking about when you were doing this?*'
Participant selects an option to a question	'*I note that you have just selected* [option A/B/C] *for item* [insert number].'	'*Can you remember why you selected this option?*'

metacognitive decision-making, they are better positioned to explain why they selected an option and which parts of the text provided them with the answer. This qualitative data provides valuable contextual data to be able to infer specific levels of processing in relation to individual TOEFL iBT items. In total, the eye-tracking session plus stimulated recall interview lasted approximately one hour per participant.

Findings

The test results from the 14 participants are displayed in Appendix 1. The maximum possible score that a participant could receive was 14. Appendix 1 also contains task data providing an indication of the relative difficulty. Item 9 was the most difficult, with only six correct responses. Conversely, Item 5 was easiest, as all participants responded correctly. This information was used to select participants and items for subsequent eye-tracking data analysis. As clarified in the section 'Methods of data analysis', six items (2, 3, 6, 9, 11 and 12) were selected for closer analysis. As the analysis also required correct responses to these six items, participants had to have at least one correct response to a basic comprehension task and at least one correct response to an inferencing task. All but one of the test-takers met this requirement. Test-taker 13 received an overall score of 7 and did not respond

correctly to any inferencing items. Her data was therefore not used in the analysis. Additionally, test-takers 2, 9 and 10 did not meet the pre-established 70% threshold for successful capturing of eye-gaze data, meaning their data was also not included. This was likely due to a combination of participants wearing eyeglasses and sub-optimal lighting conditions, which can affect the ability of the tracker to identify participants' pupils. Eye-tracking data analysis proceeded with 10 of the original 14 participants. Appendix 2 shows which items the 10 participants got correct and incorrect. Eye-tracking data was analysed for those participants who responded correctly. This meant that there were unequal numbers of correct responses for each item. Eye-tracking data was therefore weighted depending on the number of correct responses to each item. This is described in the next section.

Eye-tracking data findings

As outlined in the methodology, text and questions were specified as areas of interest (AOIs). The data presented in Table 5 and for the data analysis below is only for text engagement. Initial descriptive data for saccades, backward sweeps and fixations is provided in Table 5 according to item reading purpose for participants who responded correctly to those items (see Appendix 2).

Median numbers of fixations, saccades and backward sweeps were generally higher for participants successfully completing inferencing items than basic comprehension items. Three Wilcoxon signed rank tests were performed to compare the data, using reading purpose as a grouping variable. Average numbers of fixations, backward sweeps and saccades were obtained for each participant by dividing the total numbers by the number of items they responded to correctly. For example, the total number of saccades, fixations and backward sweeps for Participant 1 elicited by basic comprehension items were divided by three, as this participant responded to all three basic comprehension items correctly. This data is presented for

Table 5 Number of saccades, backward sweeps and fixations by reading purpose

Item	Reading purpose	Number of correct responses	Median number of forward saccades	Median number of backward sweeps	Median number of fixations	Median number of visits to text
2	Basic comprehension	8	148	51.5	205.5	3.5
6	Basic comprehension	9	72	33	107	5.5
11	Basic comprehension	7	128	49.5	179	4
3	Inferencing	9	176.5	83.5	279	3
9	Inferencing	5	115	58	190	3.5
12	Inferencing	9	132.5	56	122	1.5

all 10 participants in Appendix 3. Due to multiple significance testing, a Bonferroni-adjusted *p*-value of .02 was adopted to reduce the propensity for Type I error. The data revealed that the number of fixations for inferencing items (Md = 201, n = 10) was not significantly higher than the number of fixations for basic comprehension items (Md = 157, n = 10), $z = -1.68$, $p = .09$, $r = 0.38$. Likewise, the number of saccades for inferencing items (Md = 145, n = 10) was not significantly higher than the number of saccades for basic comprehension items (Md = 120, n = 10), $z = -2.09$, $p = .04$, $r = 0.47$. However, the number of backward sweeps for inferencing items (Md = 64, n = 10) was significantly higher than the number of backward sweeps for basic comprehension items (Md = 38, n = 10), $z = -2.39$, $p = .02$, $r = .52$ with a moderate effect size. Inferencing items display a greater number of backward sweeps relative to basic comprehension items, while having a similar number of overall fixations and saccades. The greater number of backward sweeps suggests participants demonstrated a propensity for *localised* re-reading to answer inferencing items, whereas they read a longer stretch of text fewer times for basic comprehension items. Global re-reading would result in greater numbers of fixations and saccades for inferencing items in addition to greater numbers of backward sweeps.

Median values for saccade and regression length were similar across reading purpose. The data does not indicate differences between inferencing and basic comprehension question types by saccade or regression length, which are highly consistent for all test-takers for both reading purposes (see Appendix 4). This finding further supports the interpretation of the regression count finding that test-takers were engaged in localised re-reading when completing inferencing items. Lengthier backward sweeps would indicate that they wished to re-read longer portions of text. Median fixation duration was fairly consistent across reading purpose, with four out of 10 test-takers registering greater fixation duration for basic comprehension items than for inferencing items.

Text and questions were identified as areas of interest. This allowed for the number of movements between the text and questions to be identified. This is referred to as the number of visits. A single visit indicates that test-takers only had to read the stem and options once to answer correctly. Greater numbers of visits indicate switching between the question and text in order to consider different options in relation to a developing understanding of the text (see Appendix 2). As with the data already presented, the data in Table 6 is weighted according to the number of items each test-taker responded to correctly for each reading purpose.

Most test-takers tended to visit each question between three and five times while answering both basic comprehension and inferencing items. There were some significant individual variations, for example Participants 7 and 12, who each averaged seven visits to inferencing and basic comprehension

Table 6 Weighted number of visits to the question for each reading purpose

	Number of question visits	
Test-taker	**Basic comprehension**	**Inferencing**
1	4.33	5.00
3	3.33	2.67
4	3.00	4.00
5	3.00	2.67
6	2.67	2.33
7	3.00	7.00
8	3.00	2.33
11	3.00	3.00
12	7.00	1.00
14	3.00	3.50
Median	3.00	2.83

items, respectively. A greater number of visits indicates the items which presented test-takers with difficulty. However, the number of visits was not noticeably different according to reading purpose, with the median number of visits at 3.00 for basic comprehension items and 2.83 for inferencing items.

Stimulated recall data findings

Upon completion of the test, each test-taker took part in a stimulated recall interview. Verbalisations are presented here to provide context to the eye-gaze data which suggested limited differences in item-completion procedures between inferencing and basic comprehension item types. In particular, this section presents selected stimulated recall interview data for Item 9. This item was selected because it is an inferring meaning item type and because it was the most challenging for the research participants, with only six out of 14 participants responding correctly. Item 9 also recorded the lowest number of saccades, and a relatively high number of backward sweeps, which suggest that it is a good representation of the initial finding that inferential reading items elicit careful localised re-reading. Verbalisations were considered for this item to investigate how they corresponded with each other and whether they support the finding. The explanations offered by participants were typically brief (these comments are unedited to maintain authenticity). For example, Participant 2 remarked:

> Because it says "Whigs in the northern sections of the United States."

This response was consistent with verbalisations by other participants who responded correctly. All six explicitly cited the part of the text containing the

key words 'northern parts of the country', which cohered with the wording in the fourth option of the question 'regional interests'. Participant 7 stated:

> The first half of the sentence mentions "northern" parts [of the country]. I think it should be "regional interests". No other reasons.

Nonetheless, several of the participants who responded correctly were aware of the different nature of this item type from basic comprehension and cited this as an explanation of their approach to the task. For example, Participant 5 showed awareness that the item depended more on the conceptual rather than explicit meaning of the text:

> Because it was mainly "in the Northern sections of the United States", so I thought it was "regional"… the other words [options] were in the paragraph, but the fourth one was the only one where the idea was there just because they mentioned geographical area.

Likewise, Participant 8 was also aware of this and so offered a more comprehensive explanation of her item-completion process:

> Because it said, "can be inferred from", so this [Option 3] is obviously not correct. The Whig Party is not focused on the issue of public liberty [Option 1], so this is also not correct, and I thought this [Option 4] would be more correct because in here it says in particularly the "northern sections", so I thought maybe the regional interests is correct.

Participant 11 was also aware that this task required test-takers to link meaning between the question stem, even though he was uncertain of the meaning of 'infer' in the question. When linking 'variations in political beliefs within the Whig Party' and the meaning expressed in the key, Participant 11 stated:

> I have to admit, I had no idea what "inferred" meant. But I noticed "variations in political beliefs" and it [the text] was kind of talking about in particular Whigs in the "northern sections of the United States also believed that government power should be used to foster the moral welfare of the country", so that's why I put "regional" [Option 4].

The verbalisations of successful participants are consistent with each other, and all suggest that this item targeted careful local reading. This finding is further discussed in the next section.

Discussion

Research question

The RQ which underpinned this investigation was: *What item-completion procedures do test-takers use to complete different task types in the TOEFL iBT Reading test?*

Three hypothesis statements were formulated in relation to the eye-tracking data. These were that there would be a difference in the total number of saccades, backward sweeps and fixations across items which are designed to assess either basic comprehension or inferencing. The hypothesis statements reflected the position that increased cognitive load associated with reading purpose would influence test-takers' observable engagement with the text. Interview data was used to provide further insight into the eye-gaze data, to either corroborate the findings or provide wider context to the observable engagement with the text.

The data revealed no significant differences in terms of the number of fixations or saccades for inferencing items relative to basic comprehension items. However, the number of backward sweeps for inferencing items was significantly higher than for basic comprehension items. The findings suggest that reading to complete basic comprehension and inferencing items in the TOEFL iBT both represent a form of *careful local reading* rather than expeditious or global reading (Urquhart and Weir 1998), with decisions about which part of the text to read based on an expeditious word spot strategy. More persistent backtracking by participants in completing inferencing items suggests that the text processing required for these items is of greater complexity than basic comprehension items. Participants' verbalisations indicate that a substantial amount of text processing is required to establish propositional meaning across clauses. Comments show that participants associate the pronoun 'they' with 'variations in political beliefs' from the question stem. Additionally, this is associated with the relevant sentence containing the key word 'regional'. This is further evidence that this type of cataphoric (pronominal) referencing (Khalifa and Weir 2009:51) is a crucial component of inferential reasoning which forms a significant part of the inferential reasoning construct in the TOEFL iBT Reading test (Owen 2016).

One caveat to this study is that participants were not strictly timed when completing the test. Although this likely had an impact on their test-taking strategies, the lack of familiarity with eye tracking and the necessity of securing usable data meant that it was necessary to sacrifice some ecological validity. This emphasis on careful local reading for both reading purposes is further reinforced by consistent data patterns for fixation duration (ms), saccade and regression length (pixels) and the number of

visits to the question, suggesting consistent item-completion procedures for both reading purposes. These findings cohere with those of Bax (2013), who showed that successful test-takers exhibit a tendency to identify and pay attention to key parts of the text, reading these intently. This research provides evidence that this pattern is repeated across test items regardless of whether the reading purpose is 'basic comprehension' or 'inferencing'. A second caveat is the small sample size. Although not significant, the number of fixations and saccades approached statistical significance (n=10), meaning that it is possible that a repeat of this experiment with a larger sample size may result in significant findings for fixations, saccades and backward sweeps.

Conclusions and recommendations

This study has demonstrated that eye-tracking data has some utility in providing validity evidence for claims that different item types assess inferencing or basic comprehension. The heavier cognitive load associated with items which assess inferencing, which is defined by ETS as 'the ability to comprehend ideas or connections between propositions which are not explicitly stated in the text' (Pearlman 2008:242), stimulate a greater number of backward sweeps within the text, although not a greater number of fixations overall. This suggests that inferencing as operationalised in the TOEFL iBT Reading test is accessed through a combination of expeditious word spotting and careful local reading. However, other metrics employed in this research (saccades and fixation length, number of question visits) did not indicate a difference according to reading purposes. The research also demonstrated the continuing importance of stimulated recall interviews to make stronger claims about how specific reading purposes have been operationalised by test developers.

The research revealed differences in behaviour between participants who responded correctly to the same item, consistent with McCray and Brunfaut (2018) who also highlighted this artefact of eye-tracking methodology. Eye-tracking studies to date appear to support the findings of Buck (1990, 1991), who used verbal protocol analysis to reveal that test-takers often employ highly individualised item-completion processes. This problem is of particular relevance for eye-tracking studies. During eye-tracking studies, large amounts of eye-movement (vector) and fixation data are collected from each research participant. This increases the reliability of aggregated data but does not alter the sample size, which is typically small, due to the technicalities and expense of eye tracking. It's therefore crucial to account for individual differences through the use of statistical techniques such as mixed-effects models or repeat measures ANOVA (Zuur, Ieno, Walker, Saveliev and Smith 2009).

Individualised response patterns have significant implications for defining psychological constructs in language tests. Language processing involves multiple inter-related skills, and items designed to operationalise one sub-skill will likely elicit other sub-skills. As a result, expert judges sometimes fail to agree on what individual items are assessing (Buck and Tatsuoka 1998). The risk of disagreement increases as the number of sub-constructs increase (Alderson and Lukmani 1989). Therefore, eye-tracking technology offers much promise but also many potential pitfalls for validation research in reading tests. The finely grained data makes visible previously unseen task completion processes. However, if data is too individualised, it may make generalisations about constructs difficult to establish. This chapter concludes by offering a discussion of potential pitfalls of eye-tracking methodology and some recommendations.

Reflections on eye-tracking methodology

This section outlines four 'problems' with the use of eye-tracking technology in contemporary language testing and offers some possible ways forward.

The first issue is *domain relevance*. Cognitive validity, as discussed at the beginning of this chapter, refers to the similarity of behaviour in test tasks to real-world domain-relevant behaviour. However, Bax (2013:446) describes language tests as purposefully *disruptive* forms of reading, in which readers are required to repeatedly move between the text and test task, such as a multiple-choice question. Reading tests therefore differ from 'natural' reading in an academic domain (Bax, 2013, Rayner, Pollatsek, Ashby and Clifton 2012:221). The implicit argument made by Bax is that test-taking strategies made observable through eye tracking do not reflect the kinds of reading behaviours that would be observable if eye tracking were undertaken during real-world reading in a university library. Reading in an academic domain can also be a form of disrupted reading. It is rarely a linear process. Participants may engage with multiple texts and make and revise notes while reading. Eye-tracking technology offers significant opportunities to explore the real-world construct of reading for academic purposes in more depth which could be a powerful means of influencing future test design. Eye-tracking data from real-world reading could then be used in future eye-tracking studies to compare test-elicited behaviour with domain behaviour. Such studies should always be accompanied by stimulated recall interviews which provide invaluable contextual data to interpret eye-tracking data.

The second problem is that of *identifying higher-order processing*. Models of cognitive validity in reading tests such as those offered by Khalifa and Weir (2009) and Koda (2005) include complex, higher-level processes such as inferential reasoning or building a mental representation of a text. From a validity point of view, this is a challenging statement. Given that we lack

sufficient knowledge of what higher-order reading processes look like during real-world reading, it is difficult to know what evidence we should look for during eye tracking of reading tests. Observable metrics such as backward sweeps, fixations, backtracking and saccades offer no direct evidence of processing above the lexical level (Bax 2013:445). Evidence for the efficacy of eye tracking to reveal and distinguish between higher-level processes has so far only been implied by the absence (or limited duration) of fixations on specific lexical items (Bax 2013). From a cognitive validity point of view, research designs such as those offered by Bax and Weir (2012) and Aryadoust and Ang (2019) would seem to be the most promising, as they compared response patterns across items which claim to elicit differential behaviour patterns to see what differences emerge, if any. Further studies which explore behaviour during testing and compare it to reading behaviour in the domain of interest are essential in order to further investigate reading and to determine whether key eye-tracking metrics can offer operational definitions of higher-order processing which can be investigated in reading tests.

The third problem is *individual differences* in datasets. Buck (1990, 1991) and Buck and Tatsuoka (1998) revealed that test-takers often employ highly individualised item-completion processes. As a result, there is always a possibility that eye-tracking data will differ more widely between participants in the same experimental group than between participants in different experimental groups. Thus, standard deviations or inter-quartile ranges in each group could be sufficiently wide as to undermine significance testing. If test-takers respond correctly to an item using radically different test-taking strategies, this may be an indication that our definition of the construct (or domain) is incomplete. Additionally, studies exploring individual differences in second language readers have also shown that the first language of the learners has an effect on their perceptual span (whole word and gap processing). For example, learners whose first languages are read from right to left (e.g., Arabic, Hebrew and Urdu) have shown a bias to text left of fixations, making their reading less efficient than learners used to reading from left to right (e.g., Pollatsek, Bolozky, Well and Rayner 1981, Jordan et al 2014, Paterson et al 2014). Much stricter control of participant recruitment, controlling for variables such as first language and English language proficiency, is required, although this may have an impact on study sample sizes.

The final problem to be addressed here is the issue of *replication*. The language testing studies cited in this chapter typically provide extremely limited information about the methods of data cleaning and analysis, going from descriptions of data collection to statistical hypothesis testing. In many eye-tracking software packages, saccade and regression/backward sweep data are not provided (e.g., Tobii Studio 2016). As a result, researchers must calculate saccade and backward sweep length using fixation and vector data,

as outlined in this chapter. There is no discussion of how this is achieved, how these methods compare across studies or how they are affected by the hardware used in the studies. For example, if screens have different resolutions, then data obtained from one participant could be radically different if they completed the same activity on a different computer. Likewise, if ecological validity is not preserved due to altered text size, line returns may be at different locations within sentences and paragraphs, meaning fixation, saccade and return sweep data could also be dramatically different. Given that technical requirements of eye-tracking studies usually result in small sample sizes, this hampers our ability to replicate experiments or combine datasets to create larger sample sizes which would strengthen findings. Future eye-tracking studies should seek to be international, collaborative and cross-sectional research projects, in which data collection occurs concurrently in multiple research sites with different groups of test-takers with different L1s. This would enable researchers to compare findings between groups of test-takers and present findings with larger sample sizes. Research of this kind should also seek to publicise how they standardised data collection and analytical procedures across research locations so that findings can be compared. This would provide an opportunity for larger datasets to be constructed from which multiple research questions could potentially be answered.

References

Alderson, J C and Lukmani, Y (1989) Cognition and reading: Cognitive levels as embodied in test questions, *Reading in a Foreign Language* 5 (2), 253–270.

Anderson, N J (1991) Individual differences in strategy use in second language reading and testing, *Modern Language Journal* 75 (4), 460–472.

American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME] and Joint Committee on Standards for Educational and Psychological Testing (2014) *Standards for Educational and Psychological Testing: National Council on Measurement in Education*, Washington, D.C.: American Educational Research Association.

Anderson, N J, Bachman, L F, Perkins, K and Cohen, A D (1991) An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources, *Language Testing* 8 (1), 41–66.

Aryadoust, V and Ang, B H (2019) Exploring the frontiers of eye tracking research in language studies: a novel co-citation scientometric review, *Computer Assisted Language Learning*, 1–36.

Ballard, L (2017) *The Effects of Primacy on Rater Cognition: An Eye-Tracking Study*, unpublished PhD thesis, Michigan State University.

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

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

Bereiter, C and Bird, M (1985) Use of thinking aloud in identification and teaching of reading comprehension strategies, *Cognition and Instruction* 2 (2), 131–156.

Block, E (1986) The comprehension strategies of second language readers, *TESOL Quarterly* 20 (3), 463–494.

Bowles, M A (2010) *The Think-aloud Controversy in Second Language Research*, New York: Routledge.

Brunfaut, T (2016) *Looking into Reading II: A Follow-up Study on Test-takers' Cognitive Processes While Completing Aptis B1 Reading Tasks*, British Council Validation Series, Volume VS/2016/001, London: The British Council.

Brunfaut, T and McCray, G (2015) *Looking into test-takers' cognitive processes whilst completing reading tasks: A mixed-method eye-tracking and stimulated recall study*, ARAGs Research Reports Online, Volume AR/2015/001, London: British Council, available online: www.britishcouncil.org/sites/default/files/brunfaut-and-mccray-report_final.pdf

Buck, G (1990) *The testing of second language listening comprehension*, unpublished PhD thesis, Lancaster University.

Buck, G (1991) The testing of listening comprehension: an introspective study, *Language Testing* 8 (1), 67–91.

Buck, G and Tatsuoka, K (1998) Application of the rule-space procedure to language testing: examining attributes of a free response listening test, *Language Testing* 15 (2), 119–157.

Cohen, A D and Upton, T A (2006) *Strategies in Responding to the New TOEFL Reading Tasks*, TOEFL Monograph Series Report No. 33, Princeton: Educational Testing Service.

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

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

Drewes J, Zhu W, Hu Y and Hu X (2014) Smaller is better: Drift in gaze measurements due to pupil dynamics, *PLOS ONE* 9 (10), available online: doi.org/10.1371/journal.pone.0111197

Educational Testing Service (2009) *The Official Guide to the TOEFL Test* (Third edition), Princeton: Educational Testing Service.

Gass, S and Mackey, A (2000) *A Stimulated Recall Methodology in Second Language Research*, New York: Routledge.

Holmqvist, K, Nyström, M, Andersson, R, Dewhurst, R, Jarodzka, H and Van de Weijer, J (2011) *Eye Tracking: A Comprehensive Guide to Methods and Measures*, Oxford: Oxford University Press.

Jamieson, J M, Eignor, D, Grabe, W and Kunnan, A J (2008) Frameworks for a new TOEFL, in Chappelle, C, Enright, M K and Jamieson, J M (Eds) *Building a Validity Argument for the Test of English as a Foreign Language*, New York: Routledge, 55–95.

Jordan, T J, Almabruk, A, Gadalla, E M, McGowan, V A, White, S J, Abedipour, L and Paterson, K B (2014) Reading Direction and the Central Perceptual Span: Evidence from Arabic and English, *Psychonomic Bulletin & Review* 21 (2), 505–511.

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

Koda, K (2005) *Insights into Second Language Reading: A Cross-linguistic Approach*, Cambridge: Cambridge University Press.

Komogortsev, O, Gobert, D V and Jayarathna, S (2010) Standardization of Automated Analyses of Oculomotor Fixation and Saccadic Behaviors, *IEEE Transactions on Biomedical Engineering* 57 (11), 2,635–2,645.

McCray, G and Brunfaut, T (2018) Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking, *Language Testing* 35 (1), 1–23.

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

Owen, N (2016) *An evidence-centred approach to reverse engineering: Comparative analysis of IELTS and TOEFL iBT reading sections,* unpublished PhD thesis, University of Leicester.

Paterson, K B, McGowan, V A, White, S J, Malik, S, Abedipour, L and Jordan, T R (2014) Reading Direction and the Central Perceptual Span in Urdu and English, *PLOS ONE* 9 (2), available online: doi.org/10.1371/journal.pone.0088358

Pearlman, M (2008) Finalizing the test blueprint, in Chappelle, C, Enright, M K and Jamieson, J M (Eds) *Building a Validity Argument for the Test of English as a Foreign Language*, New York: Routledge, 227–258.

Pollatsek, A, Bolozky, S, Well, A D and Rayner, K (1981) Asymmetries in the Perceptual Span for Israeli Readers, *Brain and Language* 14 (1), 174–180.

Poole, A and Ball, L (2005) Eye tracking in human-computer interaction and usability research: Current status and future prospects, in Ghaoui, C (Ed) *Encyclopedia of Human Computer Interaction*, Pennsylvania: Idea Group, 211–219.

Rayner, K (1978) Eye movements in reading and information processing, *Psychological Bulletin* 85 (3), 618–660.

Rayner, K (1998) Eye movements in reading and information processing: 20 years of research, *Psychological Bulletin* 124 (3), 372–422.

Rayner, K, Pollatsek, A, Ashby, J and Clifton, C (2012) *The Psychology of Reading*, New York: Psychology Press.

Sibley, C, Foroughi, C K, Olson, T, Moclaire, C and Coyne, J T (2017) Practical considerations for low-cost eye tracking: An analysis of data loss and presentation of a solution, in Schmorrow, D and Fidopiastis, C (Eds) *Augmented Cognition, Neurocognition and Machine Learning*, 11th International Conference, AC 2017, New York: Springer Publishing Company, 236–250.

Singer, M (2007) Inference processing in discourse comprehension, in Gaskell, M G (Ed) *The Oxford Handbook of Psycholinguistics*, Oxford: Oxford University Press, 343–360.

Stickler, U and Shi, L (2017) Eye movements of online Chinese learners, *CALICO Journal* 32 (1), 52–81.

Stickler, U, Smith, B and Shi, L (2016) Using eye-tracking technology to explore online learner interactions, in Caws, C and Hamel, M J (Eds) *Language-Learner Computer Interactions: Theory, Methodology and CALL Applications*, Amsterdam/Philadelphia: John Benjamins Publishing Company, 163–186.

Suvorov, R (2015) *Interacting with Visuals in L2 Listening Tests: An Eye-tracking Study*, ARAGs Research Reports Online, Volume AR-A/2015/1, London: British Council.

Tobii Studio (2016) *User's Manual Version 3.4.5*, available online: www.staff.universiteitleiden.nl/binaries/content/assets/sociale-wetenschappen/faculteitsbureau/solo/research-support-website/software/tobii-pro-studio-user-manual_3.4.5_08082019.pdf

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

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

Winke, P M, Gass, S and Sydorenko, T (2013) Factors influencing the use of captions by foreign language learners: An eye-tracking study, *Modern Language Journal* 97 (1), 254–275.

Yarbus, L (1967) *Eye Movements and Vision*, New York: Plenum Press.

Zuur, A F, Ieno, E N, Walker, N J, Saveliev, A A and Smith, G M (2009) *Mixed effects models and extensions in ecology with R*, New York: Springer.

Appendix 1: Test results for all participants and all items

		Item															Participant score
		1	2	3	4	5	6	7	8	9	10	11	12	13 (three items)			
Participant	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	11
	2	1	0	1	0	1	0	1	1	1	1	1	1	1	1	0	10
	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	14
	4	1	1	1	1	1	1	0	1	0	0	1	1	1	1	1	11
	5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	14
	6	1	0	1	1	1	1	1	1	0	0	1	1	1	1	1	11
	7	1	1	0	1	1	1	1	0	1	0	0	0	1	1	1	9
	8	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	12
	9	0	1	1	1	1	0	1	1	0	1	1	0	0	1	1	9
	10	1	0	1	1	1	0	1	1	0	1	1	0	1	1	1	10
	11	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	12
	12	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	12
	13	0	0	0	1	1	0	1	0	0	1	1	0	1	1	1	7
	14	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	13
Item total		11	9	12	13	14	9	13	12	6	11	11	10	13	13	13	M = 12.1

Appendix 2: Test data for participants who responded correctly to at least one basic comprehension and one inferencing item

Participant	Basic comprehension			Total	Inferencing			Total
	2	6	11		3	9	12	
1	1	1	1	3	1	0	1	2
3	1	1	1	3	1	1	1	3
4	1	1	1	3	1	0	1	2
5	1	1	1	3	1	1	1	3
6	0	1	1	2	1	0	1	2
7	1	1	0	2	0	1	0	1
8	1	1	0	2	1	1	1	3
11	0	0	1	1	1	1	1	3
12	1	1	0	2	1	0	1	2
14	1	1	1	3	1	0	1	2
Total	8	9	7	Total	9	5	9	

Appendix 3: Weighted data for number of fixations, saccades and backwards sweeps for participants who responded correctly to at least one basic comprehension and one inferencing item

	Number of fixations		Number of saccades		Number of backward sweeps	
Test-taker	**Basic comprehension**	**Inferencing**	**Basic comprehension**	**Inferencing**	**Basic comprehension**	**Inferencing**
1	231	328	182	254	53	86
3	104	201	79	140	29	72
4	187	153	133	122	57	60
5	159	205	111	150	37	48
6	139	144	93	92	36	43
7	181	405	130	293	39	103
8	91	200	52	128	31	70
11	132	181	106	163	33	63
12	251	94	159	134	69	59
14	155	215	140	163	59	65
Median	157	201	120	145	38	64

Appendix 4: Weighted data for saccade length, regression length and fixation duration for participants who responded correctly to at least one basic comprehension and one inferencing item

	Saccade length (pixels)		Regression length (pixels)		Fixation duration (ms)	
Participant	**Basic comprehension**	**Inferencing**	**Basic comprehension**	**Inferencing**	**Basic comprehension**	**Inferencing**
1	64	62	64.5	70	183	175
3	68	68	65	63.5	233	209
4	70	67	78	87.5	225	221.5
5	69	63	66.5	64	217	218
6	65	59	49	64	216	266
7	49	44	75	46	175	242
8	60	66	56.5	56	143	221
11	59.5	59	62.5	63	179.5	217
12	50	46.5	54.5	41	183	171
14	65	70.5	59	79	133	142
Median	64.5	62.5	63.5	63.75	183	217.5

6 Investigating EFL learners' cognitive processes of completing integrated writing tasks

Mikako Nishikawa
Nagasaki University, Japan

Guoxing Yu
University of Bristol, UK

Abstract

The study reported in this chapter explored the cognitive processes of completing integrated writing tasks among learners of English as a foreign language in Japan. The tasks were designed according to the specifications used by the Test of English for Academic Purposes (TEAP), which employ multiple texts and graphs to assess high school students' English writing proficiency for university admission. This mixed methods study employed a sequential explanatory design using eye tracking, questionnaires, and focus group discussions as research instruments for data collection. It explored the relationship between participants' English language proficiency and their cognitive processes involved in completing the integrated writing tasks. In particular, we examined the correlations between 42 participants' eye movements and their TEAP Writing sub-scores in five assessment areas (Main Idea, Coherence, Cohesion, Lexical Range and Accuracy, and Grammatical Range and Accuracy). Data from the questionnaire and the focus group discussions, in which 24 participants took part, provided further evidence on the differences between the upper and lower-proficiency groups in summarising the trends described in the texts and graphs. Overall, the findings of the study suggest that the participants' English language proficiency affected their usage of the source texts and graph information in integrated writing. Implications of the findings are discussed with reference to the effects of source input on test-taking processes and test performance, as well as the use of eye tracking as a method to examine the cognitive validity of integrated writing tasks.

Introduction

Integrated writing tasks have gained popularity in the field of language assessment because of their authenticity. While an independent writing task assesses test-takers' abilities to write an essay based on their prior knowledge of a topic or recollection of past experience, integrated writing tasks assess test-takers' abilities to report and synthesise the content, facts, and ideas from visual inputs and source texts. Integrated writing tasks are commonly seen in language assessment for academic purposes. For example, TOEFL iBT® has integrated writing tasks based on written and spoken stimuli. IELTS uses visuals such as graphs in the prompts in its academic writing tasks. The basis for our study was the Test of English for Academic Purposes (TEAP), which uses integrated tasks to assess test-takers' writing proficiency in Japan. TEAP was launched in 2014 to assess high school graduates' academic readiness in terms of English language proficiency for studying at a university. The test design is guided by Weir's (2005) socio-cognitive processing model, which emphasises both language use and the cognitive processes underlying task performance, and the test tasks frequently employ source texts and graphs as prompts. The TEAP Writing module has two writing tasks: Task A is a summary essay; Task B is an argumentative essay based on two types of graphs (e.g., a bar graph and a pie chart) and two source texts (e.g., a newspaper article and a letter-to-the-editor). Test-takers need to address the main points in the source texts and the overall trends shown in the graphs and then state their opinions in 200 words or more within approximately 40 minutes. Our focus in this chapter is Task B.

Integrated writing using source texts is arguably one of the most important skills for academic writing. Previous studies have found that the quality of source-based writings and the writing processes may be affected by a range of factors including the features of source texts (e.g., Yu 2008) and test-takers' familiarity with the discourse types of the source texts (e.g., Delaney 2008, Yu 2009). A considerable increase in using integrated writing tasks is foreseeable in university admission tests, such as TEAP, since they are believed to better distinguish between novice and advanced writers in an academic setting (Plakans 2008). However, our understanding of the integrated writing construct is still limited. Only a small body of research has addressed the effects of the features in the source input on integrated writing performance. By exploring the cognitive processes involved in completing an integrated writing task through a mixed methods approach, this chapter attempts to shed light on the use of multiple source texts and graphs in writing assessment.

Literature review

Cognitive processes of integrated writing

The research on the processes of second language writing started about 50 years ago in the early 1970s, shifting its focus from 'writing products' to 'writing processes' (Grabe and Kaplan 1996). One of the most influential studies focusing on writing processes was reported by Emig (1971, 1983) using 'verbal protocol analysis' (Grabe and Kaplan 1996:90). This study triggered a trend towards using think-aloud protocols to understand the writing processes, including conceptualising a seminal cognitive model of writing (Flower and Hayes 1981) that comprises task environment, the writers' working memory, and multiple cognitive processes (Weigle 2002).

In Flower and Hayes' model, there are two types of meta-cognitive processes involved in writing: 'knowledge telling' and 'knowledge transforming'. According to Weigle (2002), the findings from the study conducted by Scardamalia and Bereiter (1987) suggest that writing based on knowledge telling is similar to a spontaneous conversation in which little planning and revision is made. On the other hand, knowledge transforming involves writers' more skilful efforts to create new insights for their audience. In other words, writers construct meaning as they compose texts by interpreting texts that they read (Spivey 1990). In this sense, integrated writing requires more knowledge transforming than knowledge telling skills.

Another feature of Flower and Hayes' model (1981) is their three essential cognitive processes in writing: planning, translating (i.e., composing), and reviewing. Most noticeably, their model implies that the writing process is a 'hierarchical process' embedded with 'recursion' (1981:375), rather than a linear process. Similarly, Grabe and Kaplan (1996) suggest that writing requires the cognitive processes of planning, using existing knowledge, considering audience interests, addressing purposes, and staying cohesive and logical throughout the text in the absence of spontaneous feedback. In addition to these cognitive processes, it is important to include reading source texts in integrated writing.

Chan (2013) examined the similarities and differences of the cognitive processes elicited by integrated writing in real-life academic writing and in a writing test. Her survey data showed that high- and low-achieving groups behaved significantly differently in their 'task representation' (i.e., how test-takers interpret the tasks), 'selecting relevant ideas', 'organizing ideas', and 'monitoring and revising' (2013:213). Based on correlational analysis, she found that the participants' cognitive processes involved in real-life academic writing tasks were similar to those involved in reading-into-writing test tasks. Her findings based on exploratory factor analysis also suggest that

candidates' performance on reading-into-writing tasks can predict their performance in real-life academic writing tasks.

Cognitive processes of graph-based writing

Recent works that attempted to validate integrated writing tasks used different kinds of inputs. Graph-based writing was extensively investigated. Most of the previous studies focused on the process of interpreting the graphs (e.g., Carswell, Emery and Lonon 1993). More recent studies examined the effects of test-takers' graph familiarity, and the use of the graphic information in the writing output (e.g., Yang 2012, Yu and Lin 2014, Yu, He and Isaacs 2017, Yu, Rea-Dickins and Kiely 2011). Yu et al (2011, 2017) studied test-takers' cognitive processes while doing the graph-based IELTS Academic Writing Task 1, using think-aloud protocols and eye tracking, respectively. The two studies had very similar findings and concluded that factors including the types of graphs, test-takers' graphicacy (knowledge of and skills in using graphs), and writing proficiency affected both writing performance and writing processes.

Yu et al (2011) found that the type of graph (e.g., line graph, bar graph) affected the written products, as evidenced in the use of different words in three tasks each using a different type of graph. The level of familiarity with graphic conventions influenced the way that writers processed the graphic information. Similar findings were observed in Yu et al (2017), who used eye tracking for data collection. The eye-movement data showed that on average the participants spent less than 10% of the time reading the task instructions, 20% reading graphs and 70% writing. The visualisations of the eye movements illustrated each participant's test-taking processes. The impacts of graph features on the writing process were mainly observed on the aggregated metrics of eye movements in terms of total fixation duration and total visit duration. The participants had different patterns of engagement with different types of graphs, which also affected their judgement about the difficulty of each graph type. Although graph familiarity did influence the way that the participants processed the graphs, such effects were found to be 'weak and short-lived' (2017:4). They also found that the correlations between the participants' English writing ability and their eye-movement metrics were rather weak and fuzzy. The findings of Yu et al (2011, 2017) offered some glimpses into the complex nature of the IELTS graph-based writing tasks, and the dynamic interplays between test-taker characteristics (e.g., graph familiarity, English writing ability) and task features (e.g., different types of graph, the amount of information contained in a graph).

Yu and Lin (2014) compared test-takers' cognitive processes of completing IELTS and General English Proficiency Test (GEPT) graph-based writing tasks. In both tasks, test-takers need to summarise the

graphic information, but GEPT also requires test-takers to write about their personal interpretations of the phenomenon described in the graphs. In contrast, IELTS test-takers are penalised for doing so. Similar to Yu et al (2011, 2017), this study found differential impacts of graph prompts, test-takers' graphicacy and writing ability on the cognitive processes involved in writing. The differences in the participants' cognitive processes when completing the two tasks were clearly observed, especially in the second part of the GEPT tasks when test-takers started to write about their personal interpretations of the graphic data. It was also found that the type of graph had almost negligible impact on the participants' test performances and that their performances in IELTS and GEPT integrated writing tasks were highly correlated.

Eye tracking as a research method

This study employed eye tracking as the main data collection method to explore how test-takers incorporate information from multiple source texts and graphs in a TEAP task. Traditionally, research into underlying constructs in language assessment relies on think-aloud protocols or stimulated recall interviews (e.g., Yu et al 2011, Yu and Lin 2014). While think-aloud protocols are useful to reveal the cognitive processes during writing, data elicited by the technique may be constrained by the participants' short-term memory capacity, which requires talking and processing information simultaneously (Plakans 2008). Also, some studies reported that participants could unintentionally self-correct their behaviours while thinking aloud (e.g., Bridges 2010, Plakans 2009a).

Eye-tracking technology is becoming increasingly accessible to researchers. Most often, eye tracking is used in combination with stimulated-recall interviews; as Brunfaut and McCray (2015) argue, eye-tracking data cannot independently explain certain behaviours and it is always useful to have self-reported data to make sense of what is happening in the recordings of eye movements. For example, Suvorov (2015) used an eye-tracking device to compare audio and video contents when studying the construct of video-based L2 listening assessment. Bax (2013a, 2013b) investigated 38 participants' eye-movement data to identify the distinctive behaviours of successful and unsuccessful readers, by measuring their cognitive and metacognitive processes during reading. Brunfaut and McCray (2015:17) suggested that eye tracking could be a useful tool to measure global processing (i.e., summative measures of task completion), text processing (i.e., specific to the text of the item) and task processing (i.e., the interactions between the text and the response options) as a means of differentiating participants at different proficiency levels.

Research methods

Research aims and questions

This study aims to investigate the cognitive processes of Japanese learners of English when they write in response to the TEAP integrated writing tasks that use multiple graphs and source texts as prompts (see Nishikawa 2018). Our overarching research aim is to explore the relationship between the participants' English language proficiency and their cognitive processes involved in completing the integrated TEAP Writing tasks, with two research questions (RQs):

RQ1: To what extent is learners' TEAP Writing performance, as shown in the five sub-scores, associated with their eye movements?
RQ2: To what extent do learners think their English language proficiency affects their cognitive processes in completing integrated writing tasks?

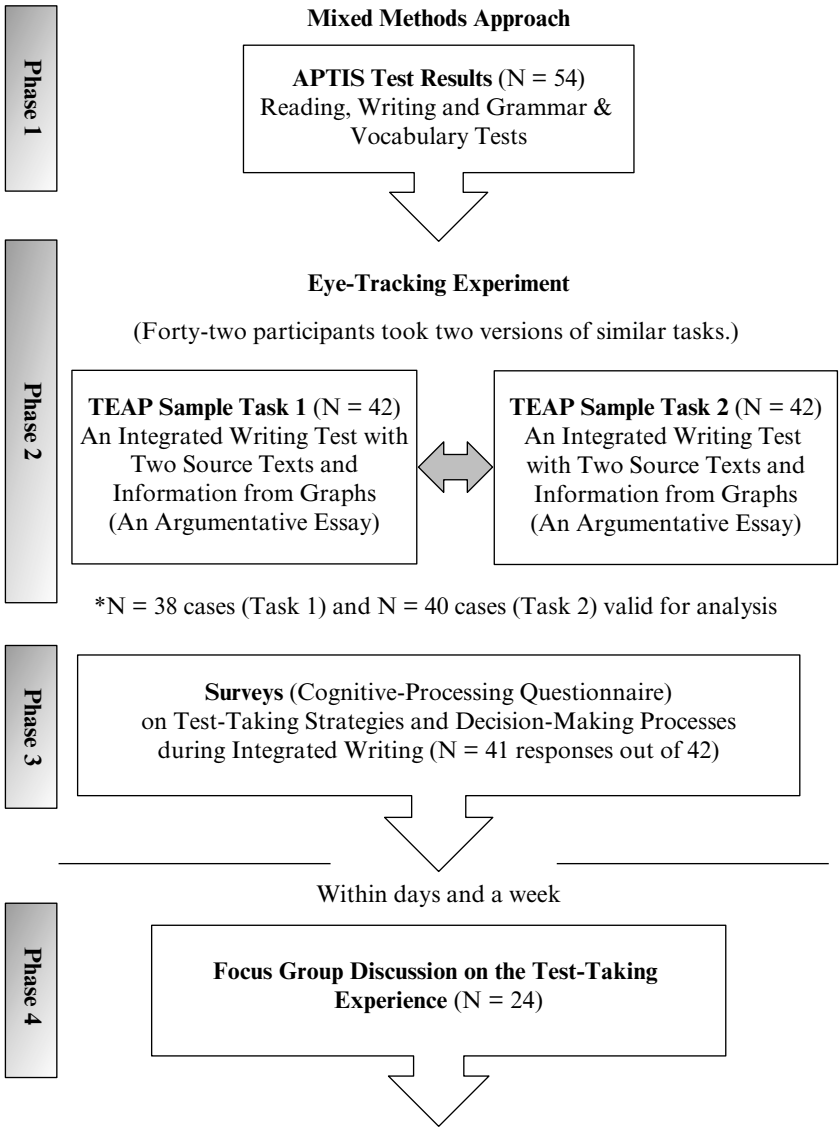
This mixed methods study employed a sequential explanatory research design, which consists of four phases (Figure 1). The first and second phases involved collecting and analysing quantitative data to inform the third and fourth phases of qualitative data analysis (see Nishikawa (2018) for further details).

In Phase 1 the participants' English proficiency was assessed using the Aptis test. Phase 2 comprised the eye-tracking experiments in which the participants completed two TEAP Writing tasks. Phase 3 included two surveys administered to the participants immediately after the eye-tracking experiments. Lastly, Phase 4 was a delayed focus group discussion joined by a small group of voluntary students.

Participants

Forty-two students volunteered to participate in the eye-tracking experiments from four Japanese high schools, two public and two private, from Kobe, Kyoto and Osaka prefectures, in the western part of Japan. They came from different grades: 15 from Grade 12, 19 from Grade 11, and eight from Grade 10. The participants spoke Japanese as their first language and studied English as their foreign language at school. The participants' ages ranged from 16 to 18 years old, which means that they had been studying English for at least three years in middle school. Ethical clearance was fulfilled by collecting the consent form signed by both the participants and their parents. The participants first took the Aptis Reading and Writing tests. In another session, they completed the two TEAP Writing tasks. Their eye movements when completing the TEAP Writing tasks were recorded using Tobii Studio.

Figure 1 Research procedure



The Aptis test was administered in advance to gauge the participants' English language proficiency. The reading and writing test scores of the 40 participants who successfully completed the eye-tracking experiment were 32.35 (SD = 8.66) and 37.68 (SD = 9.40) in a score scale of 50. The participants seemed to have slightly higher writing ability than reading ability.

Figure 2 Task 1

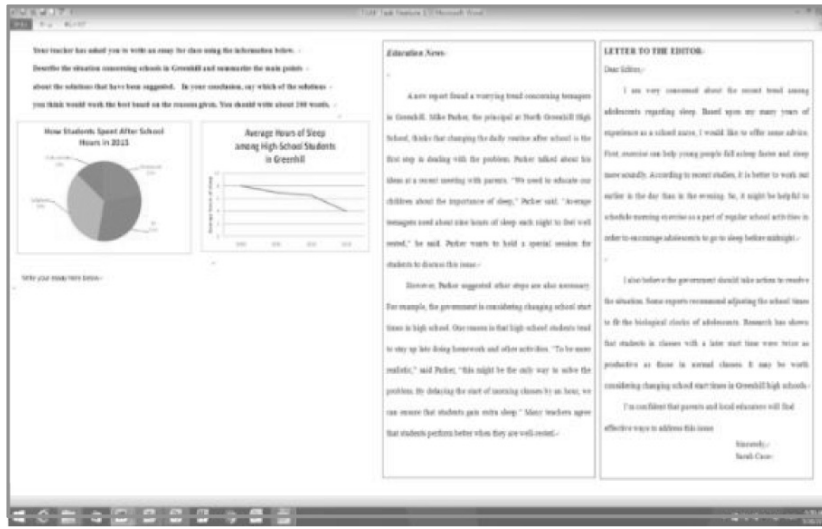
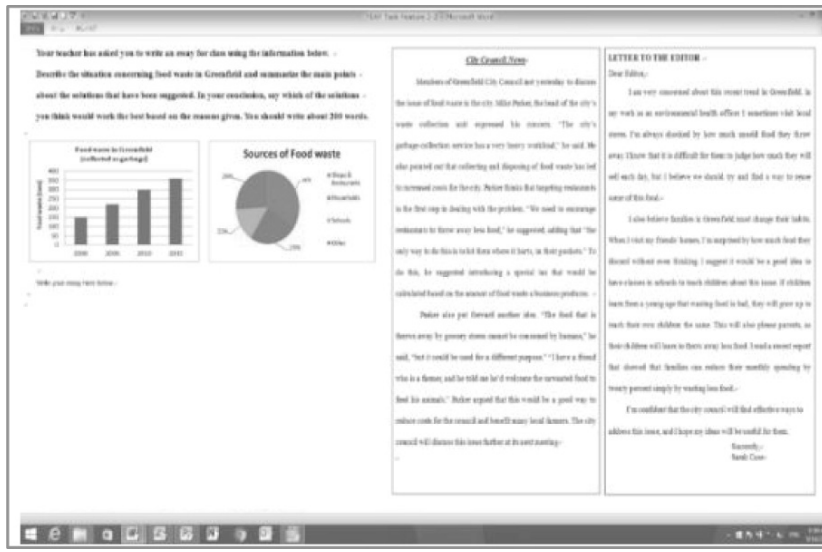


Figure 3 Task 2



The two writing tasks

Unlike the IELTS and GEPT graph-based writing tasks in our previous studies, the TEAP Writing task in the present study (see Appendix 1) includes two short texts and graphs as prompts, which is unique in at least three aspects in task design. First, it is an integrated writing task based on two related source texts, not a single source text as in many other studies. Second, it is a task including both substantial source texts and related graphs. Third, the TEAP task requires both summarisation of the source information (texts and graphs) as well as a personal interpretation of this input.

The two writing tasks used in the present study were designed according to the specifications of TEAP Writing Task B, which uses multiple source texts and graphs as task input. The tasks were presented on a computer screen for the purpose of collecting eye-tracking data (see Appendix 2).

Each task consisted of two source texts and two types of graphs. The two source texts were presented on the right half of the screen and the task instructions and graphs on the left (see Figures 2 and 3). Below the graphs, a blank space of around ¼ of the screen size was set up for the participants to enter their responses. The two tasks were made similar in terms of task instruction, length, and lexical level, according to the TEAP specifications. The two tasks were reviewed and approved by the EIKEN test developer before they were used in the study. Participants were allowed 40 minutes to complete each task, and asked to take a 10-minute break between the two tasks to reduce fatigue. Each participant did each of the two tasks in the same order. The ordering effect may be a limitation of the research. All the participants were given a paper-based sample test of TEAP a week before the study and received an orientation about the TEAP Writing task.

The essays for Task 1 and Task 2 were assessed according to five criteria: (1) Main Idea, (2) Coherence, (3) Cohesion, (4) Lexical Range and Accuracy, and (5) Grammatical Range and Accuracy. We followed the

Table 1 Correlation between test scores by the two raters

TEAP Writing scale criteria	Task 1 (n = 40)	Task 2 (n = 39)
Main Idea	.227	.516**
Coherence	.661**	.692**
Cohesion	.665**	.408*
Lexical Range and Accuracy	.509**	.475**
Grammatical Range and Accuracy	.522**	.398*
Total score	.436**	.620**

** *Correlation is significant at the 0.01 level (2-tailed)*
* *Correlation is significant at the 0.05 level (2-tailed)*

official rating scale of TEAP (www.eiken.or.jp/teap/merit/index.html) using a score scale from 0 to 8 for each criterion (see Appendix 2). Two trained raters marked the essays independently (40 essays for Task 1 and 39 essays for Task 2, respectively). Pearson correlations were computed to show the associations between the TEAP sub-scores awarded by the two raters (see Table 1). Moderate correlations were found in most marking criteria but a weak correlation was found in the Main Idea criterion in the ratings on Task 1 responses. Additionally, Rater 1 was found to be harsher than Rater 2. After finishing Task 1 ratings, the two raters met and discussed their rating criteria, which seemed to improve the inter-rater reliability in Task 2 ratings, especially in Main Idea and the overall score.

We used the average scores of the two raters to investigate the association between the TEAP sub-scores of the two tasks and eye-tracking data. The correlation was examined between the two tasks (as shown in Table 2 in 'Findings and discussion').

Eye tracker

In this study, a Tobii eye tracker TX300 with a frame rate of 300 Hz, and computer screen resolution of 1,920 × 1,080, was used. Tobii Studio was used to record eye movements and calculate eye-movement metrics in the Areas of Interest (AOIs). The AOIs are the areas within the displays where we can define and analyse the eye-movement patterns. (See how we defined AOIs in Figures 4 and 5.) Analysis of eye movements within the AOIs allowed us to make inferences about the reading and writing processes and postulate the underlying processes of integrated writing.

Figure 4 AOIs in Task 1

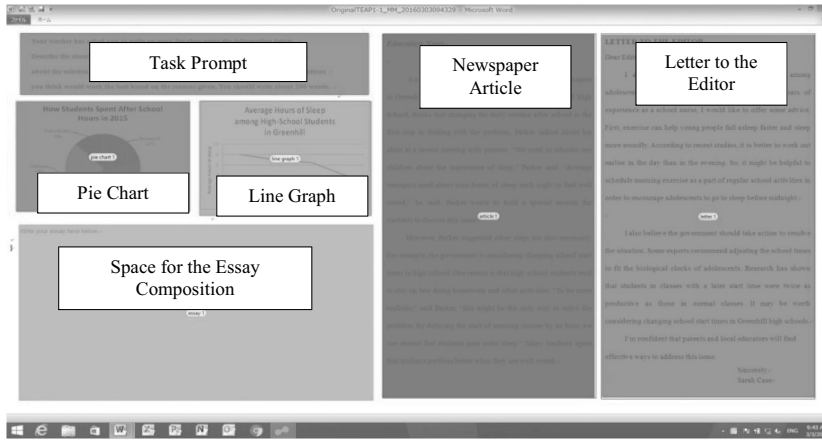
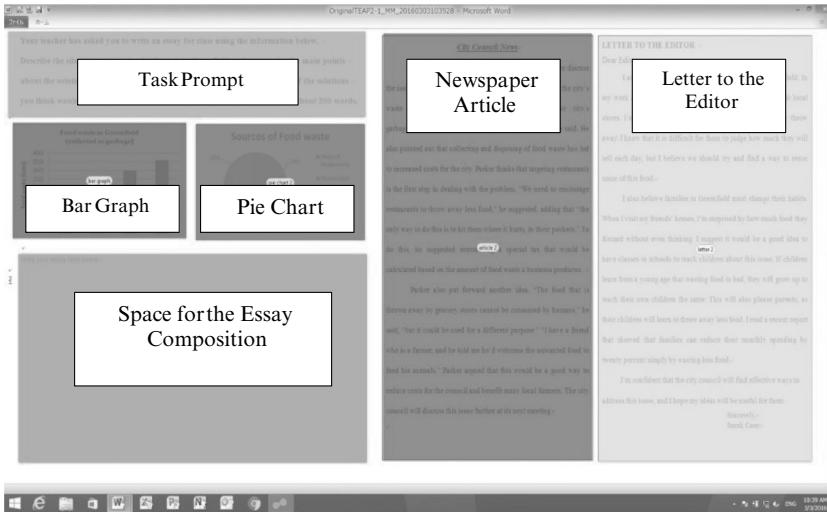


Figure 5 AOIs in Task 2



Although eye-movement data allows us to see the visual attention of the participants during a specific point in time (fixation), longer durations or frequent visits are not necessarily attributed to higher or lower English proficiency. To gain an accurate understanding of how participants engaged with source input, we decided to triangulate eye-tracking data with responses to survey and focus group discussions.

Non-parametric Spearman rank correlations were performed to identify possible associations between the participants' eye movements and their writing performances in each of the five assessment criteria (Main Idea, Coherence, Cohesion, Lexical Range and Accuracy, and Grammatical Range and Accuracy).

Cognitive processing questionnaire

The questionnaire was developed by drawing on some of the inventories used to investigate test-taking strategies in previous studies (e.g., Bridges 2010, Phakiti 2003, Xu and Wu 2012, Yang 2012). The content of the questionnaire was, however, tailored to the specific task descriptions (see Appendix 3). Since eye-tracking data alone was not sufficient to explain participants' specific decisions during test-taking, the questionnaire data helped to interpret the observations of the eye-tracking movements. Therefore, we also analysed questionnaire results by running Chi-square tests to find the differences between upper- and lower-ability groups according to their English language proficiency levels (i.e., A1 and A2 as a

lower group and B1 and B2 as an upper group measured by TEAP Writing scores).

The questionnaire and interviews provided further complementary data to understand how test-takers incorporated information from the source texts and the graphs. We categorised the responses by two groups of students, higher and lower proficiency, to explore how their English language proficiency affected their writing processes in these tasks.

Focus group discussion

We used focus group discussions rather than stimulated recall interviews with each participant, in case they misinformed interviewers by telling authorities (i.e., adults) what they wanted to hear. Focus group discussions saved time for qualitative data collection, and participants probably felt more relaxed and willing to share their honest views about their test-taking experiences in the presence of classmates.

The focus group discussion aimed to gain a deeper understanding of the cognitive processes involved in integrated writing. The semi-structured question topics included: (1) how they decided what information to include, (2) how they planned when they began writing, (3) what they often did while writing an essay, (4) what they did after they had finished an essay, and (5) which part of the TEAP Writing tasks they worried about the most (Nishikawa 2019).

A total of 24 students participated in the focus group discussions (see Appendix 4), which were facilitated by the first author across four school sites in the participants' first language (Japanese). We had planned to include all the participants in the focus group discussions but some graduated and left high school after the eye-tracking experiments. Given the circumstance, students who were available to join the focus groups sat together in a group of four in the following weeks and discussed the semi-open answers for 60 minutes each. Due to the class size (multiple groups) and time constraints, the responses were recorded only by notetaking. In order to report how they processed their reading-into-writing tasks, the notes were grouped according to the CEFR levels of the participants.

Findings and discussions

First, we compared participants' performance in five marking criteria between the proficiency groups divided on the basis of the TEAP cut scores between A2 and B1. The A2 or lower-intermediate group had 24 students and the B1 or upper-intermediate group had 16 students. Table 2 reports the Mann-Whitney U test results between the two groups.

In both Task 1 and Task 2, the upper-intermediate group outperformed the lower-intermediate group in all five TEAP sub-scores (Main Idea,

Table 2 Results of the Mann-Whitney U tests on the five sub-scores between the upper and lower groups

Ranks						
TEAP Task 1	**Two groups**	**Mean rank**	**Sum of ranks**	**U**	**Z**	**P**
Main Idea (Task 1)	Lower-intermediate	15.27	366.50	66.500	−3.493	.000
	Upper-intermediate	28.34	453.50			
Main Idea (Task 2)	Lower-intermediate	15.42	370.00	70.000	−3.420	.001
	Upper-intermediate	28.13	450.00			
Coherence (Task 1)	Lower-intermediate	16.54	397.00	97.000	−2.649	.008
	Upper-intermediate	26.44	423.00			
Coherence (Task 2)	Lower-intermediate	16.13	387.00	87.000	−2.971	.003
	Upper-intermediate	27.06	433.00			
Cohesion (Task 1)	Lower-intermediate	16.04	385.00	85.000	−2.987	.003
	Upper-intermediate	27.19	435.00			
Cohesion (Task 2)	Lower-intermediate	15.13	363.00	63.000	−3.632	.000
	Upper-intermediate	28.56	457.00			
Lexical Range (Task 1)	Lower-intermediate	17.27	414.50	114.500	−2.163	.031
	Upper-intermediate	25.34	405.50			
Lexical Range (Task 2)	Lower-intermediate	14.83	356.00	56.000	−3.840	.000
	Upper-intermediate	29.00	464.00			
Grammatical Range (Task 1)	Lower-intermediate	17.42	418.00	118.000	−2.074	.038
	Upper-intermediate	25.13	402.00			
Grammatical Range (Task 2)	Lower-intermediate	14.65	351.50	51.500	−3.939	.000
	Upper-intermediate	29.28	468.50			

Coherence, Cohesion, Lexical Range and Accuracy, Grammatical Range and Accuracy).

Research Question 1

To explore the extent to which eye movements influence writing performance, non-parametric Spearman rank correlations were calculated between the eye-movement metrics and the five sub-scores in all the AOIs.

Since all AOIs are different in size, reporting absolute eye-tracking measures would skew the results. To avoid this issue, we only report the ratio of fixation duration in AOIs in association with the five TEAP assessment criteria. The ratio of fixation duration is the sum of fixation duration of a particular AOI (e.g., sum of fixation duration in the Task Prompt) divided by the sum of fixation duration of all AOIs (e.g., Total Sum of all AOIs).

In Task 1, there were no statistically significant correlations found in any of the variables. In Task 2, however, a small-to-medium, negative correlation was found between the ratio of fixation duration in Task Prompt and Coherence ($r = -.34$, $p < .05$), as shown in bold. The result indicated that the longer the participants viewed the task prompt, the lower the coherence score they were awarded in Task 2. This is understandable since the participants already knew the task prompt well after completing Task 1 and the more able students could dedicate their time to their essay instead of re-visiting the prompt.

Table 3 Correlations between the ratio of fixation duration and the five criteria scores (Task 1)

Spearman rank correlations (Task 1), N = 40						
		Main Idea	**Coherence**	**Cohesion**	**Lexical Range and Accuracy**	**Grammatical Range and Accuracy**
Task prompt	Correlation coefficient	.255	.238	.262	.235	.232
	Sig. (2-tailed)	.112	.140	.102	.145	.149
Essay composition	Correlation coefficient	.000	.068	.123	.152	.176
	Sig. (2-tailed)	1.000	.676	.448	.348	.278
Line graph	Coefficient correlation	−.063	−.037	−.044	−.073	−.125
	Sig. (2-tailed)	.698	.821	.788	.655	.441
Pie chart	Correlation coefficient	−.120	−.107	−.088	−.133	−.188
	Sig. (2-tailed)	.459	.513	.587	.412	.244
Newspaper article	Correlation coefficient	−.179	−.127	−.193	−.112	−.153
	Sig. (2-tailed)	.270	.436	.234	.490	.347
Letter to editor	Correlation coefficient	.267	.235	.219	.283	.201
	Sig. (2-tailed)	.096	.145	.175	.077	.215

Table 4 Correlations between the ratio of fixation duration and the five criteria scores (Task 2)

Spearman rank correlations (Task 2), N = 39						
		Main Idea	**Coherence**	**Cohesion**	**Lexical Range and Accuracy**	**Grammatical Range and Accuracy**
Task prompt	Correlation coefficient	−.295	**−.335***	.056	−.192	−.110
	Sig. (2-tailed)	.065	**.034**	.732	.235	.499
Essay composition	Correlation coefficient	.124	.158	.018	.314*	.146
	Sig. (2-tailed)	.445	.331	.914	**.048**	.369
Bar graph	Correlation coefficient	−.142	−.197	−.059	−.156	−.165
	Sig. (2-tailed)	.382	.224	.719	.337	.310
Pie chart	Correlation coefficient	.090	−.187	.109	.056	.014
	Sig. (2-tailed)	.581	.247	.504	.733	.932
Newspaper article	Correlation coefficient	−.053	−.040	−.038	−.239	−.184
	Sig. (2-tailed)	.744	.805	.815	.137	.256
Letter to editor	Correlation coefficient	−.076	.132	.134	−.049	−.028
	Sig. (2-tailed)	.642	.417	.410	.766	.866

** Correlation is significant at the 0.05 level (2-tailed)*

In Task 2, a small-to-medium positive correlation ($r = .31$, $p < .05$) was found between the ratio of fixation duration in Essay Composition and Lexical Range and Accuracy, as shown in bold. The result could imply that the longer the participants spent on the screen to write their essay, the higher the score they were awarded on Lexical Range and Accuracy. The survey responses and focus group discussions confirmed that the students in the lower-proficiency group found the integrated writing task challenging due to their limited lexical knowledge. Unfortunately, the eye-movement data was inadequate to explain why a positive correlation was only found with Lexical Range and Accuracy but not with other assessment criteria. These significant correlations, however, should be interpreted with great caution due to the large number of comparisons made. It was also difficult to determine whether longer fixation duration in Essay Composition would necessarily lead to better performance in writing. Therefore, it is important that we cross-check

these findings with qualitative analysis using the questionnaire responses and focus group discussions.

Research Question 2

Since eye-movement data alone could not provide the whole picture of test-takers' cognitive processes, we examined the questionnaire and focus group data to further examine how learners' cognitive processes of integrated writing might have been affected by different features of the task. Table 5 reports the students' views of the comprehensibility of the graphic input on a 5-point Likert scale, including line graph (Task 1), pie chart (Task 1, Task 2), and bar graph (Task 2). Also, the participants reported their levels of understanding of the titles, values on the graphs, the x and y axes, and the easiness of summarising the main features of the graphs.

As reported in Nishikawa (2019), line graphs (Task 1) and bar graphs (Task 2) were viewed for longer than the pie charts (Task 1 and Task 2). The questionnaire data as shown in Table 5 revealed that the majority of the participants found it easy or very easy to understand the information in all types of graphs. Over 80% of the upper-intermediate students found that the line graph in Task 1 was neither easy nor difficult to understand while only 5% of the lower-proficiency students had the same response. The percentage of students who chose 'Difficult' was similar between the two groups. No student in the higher-proficiency group chose 'Very Difficult' in contrast to 18% in the lower-proficiency group. Likewise, most of the students considered the pie chart in Task 1 either 'Easy' or 'Very Easy' to understand. Similarly, most participants found the bar graph and pie chart in Task 2 'Easy' or 'Very Easy'. While most participants found graphic information easy to understand, weaker students tended to rate line graphs as slightly more difficult.

Through the focus group discussions, it became clearer what they meant by 'difficult'. It was difficult for them to interpret the trends in the graphs and to summarise those trends in English. The data revealed that more than half of the lower-proficiency students found summarisation in English very difficult. Lack of familiarity in English vocabulary made it harder for the participants to explain the main points even though they did understand the trends presented in the graphs. A series of Chi-square tests showed that there were no statistically significant differences in any of the questionnaire responses between the two groups of participants. This may explain the earlier finding that more capable students spent greater effort in describing the visual input than summarising the text.

The focus group discussions revealed that participants' language proficiency clearly played a role in their writing processes. For example, an upper-level student (B2) said, 'I planned my essay by summarising the

Table 5 Survey responses on understanding of the graphs

Graph survey	Very difficult	Difficult	Neutral	Easy	Very easy	N/A	Total (N = 37)
(Q #2_1)	How easy or difficult was it for you to understand the line graph in Task 1?						
Lower-intermediate	18%	14%	36%	18%	9%	5%	22 (100%)
Upper-intermediate	0%	13%	40%	7%	40%	0%	15 (100%)
(Q#2_2)	How easy or difficult was it for you to understand the pie chart in Task 1?						
Lower-intermediate	5%	0%	19%	43%	33%	0%	22 (100%)
Upper-intermediate	0%	6%	6%	38%	50%	0%	15 (100%)
(Q #2_3)	How easy or difficult was it for you to understand the bar graph in Task 2?						
Lower-intermediate	5%	0%	19%	43%	33%	0%	22 (100%)
Upper-intermediate	0%	0%	6%	50%	44%	0%	15 (100%)
(Q#2_4)	How easy or difficult was it for you to understand the pie chart in Task 2?						
Lower-intermediate	5%	0%	14%	52%	29%	0%	22 (100%)
Upper-intermediate	0%	0%	6%	44%	50%	0%	15 (100%)
(Q#5_1)	How easy or difficult was it for you to understand the titles of the graphs?						
Lower-intermediate	9%	5%	43%	33%	10%	0%	22 (100%)
Upper-intermediate	0%	12%	19%	50%	19%	0%	15 (100%)
(Q#5_2)	How easy or difficult was it for you to understand the values on the graphs?						
Lower-intermediate	0%	10%	14%	52%	24%	0%	22 (100%)
Upper-intermediate	0%	6%	12%	69%	13%	0%	15 (100%)
(Q#5_3)	How easy or difficult was it for you to understand the unit of the x and y axes on the graphs?						
Lower-intermediate	14%	0%	29%	33%	24%	0%	22 (100%)
Upper-intermediate	6%	0%	13%	56%	25%	0%	15 (100%)
(Q#5_4)	How easy or difficult was it for you to interpret the information in the graphs?						
Lower-intermediate	10%	19%	33%	33%	5%	10%	22 (100%)
Upper-intermediate	0%	6%	19%	69%	6%	0%	15 (100%)
(Q#5_5)	How easy or difficult was it for you to summarise the main trends of the graphs in English?						
Lower-intermediate	19%	48%	19%	9%	5%	19%	22 (100%)
Upper-intermediate	12%	25%	44%	19%	0%	12%	15 (100%)

graphic information and organising my thoughts in my mind.' Another upper-level student (B1) also claimed that he began by including information from graphs and his opinions about the issue. In contrast, some lower-level students (A2) indicated that they started writing without any planning.

Furthermore, Table 6 summarises the students' recollections of how they engaged with the two source texts when they summarised the key information in them.

According to the focus group data, a majority of the participants at the upper-intermediate level (44%) began composing their essays by re-reading the whole text once or twice, or by going back to specific paragraphs (31%). Approximately one-third of the participants at lower-intermediate level (33%) reported that they had decided which information to include by searching for some keywords and 10% indicated that they had relied on their memory of their own experience and had written an independent essay based on their knowledge about the topic.

It is worth noting that about a quarter of the students (24% at lower-intermediate, 37% at upper-intermediate) admitted that they had either copied the sentences directly from the source texts or borrowed some phrases. Copying a long string of words and phrases indicated their lack of proficiency in paraphrasing or summarising. The focus group discussions provided evidence that even upper-level students struggle with paraphrasing. As one upper-level student (B1) said, 'I borrowed some words that people said in the texts.'

The focus group discussions also showed that some of them did not know what they were supposed to write (despite a test orientation session

Table 6 Use of the source texts for summarisation (Questionnaire# 6–2)

Source texts	**Deciding which information to include from the source text**					
	Never thought about it	By re-reading the whole text once or twice	By going back to a specific paragraph of some importance	By scanning and looking for the keywords	By memory	Total
Lower-intermediate	5%	14%	38%	33%	10%	21 (100%)
Upper-intermediate	0%	44%	31%	19%	6%	16 (100%)
X^2 (4, N =37) =.586, p>.05						
	Connecting the ideas from the source texts					
	Never thought about it	By copying the sentences from the source texts	By borrowing some words and phrases from the source texts	By referring to ideas from the source texts	By memory	Total
Lower-intermediate	5%	24%	38%	24%	9%	22 (100%)
Upper-intermediate	6%	37%	38%	13%	6%	15 (100%)
X^2 (4, N =37) =8.12, p>.05						

offered immediately before the test). In the focus group discussions, some reported that they did not know how to borrow ideas and synthesise information. An A2 student said, 'I relied on what the main characters said in the text … I looked at the last sentence of each paragraph … After I have read the text, I just began writing.' This supports some findings in earlier studies (e.g., Plakans 2009b) that integrated writing reveals discourse synthesis and that students without experience in synthesising sources could be negatively impacted. The focus group discussions also revealed that the students with lower proficiency levels were more worried about making spelling and grammar mistakes than writing cohesive paragraphs.

Conclusion

This study investigated the extent to which test-takers' eye fixation durations in six areas of interest (two source texts, two graphs, task prompts, and the writing space of the integrated writing tasks) were correlated with the scores they received in five assessment criteria (Main Idea, Coherence, Cohesion, Lexical Range and Accuracy, and Grammatical Range and Accuracy). The only positive correlation was found between Lexical Range and Accuracy and the ratio of fixation duration in essay composition. The data collected from the questionnaire and focus group discussions helped us to get a better understanding of the complex relationships between eye movements and test scores as well as the dynamic cognitive processes involved in completing the integrated writing tasks.

From the actual writing, we noticed students in the lower-proficiency group tended not to paraphrase or summarise source information. Also, there was a tendency not to follow the instruction in the task prompt. This may suggest that the first 10 minutes of total fixation duration of the task prompt was a good indicator of a participant's reading proficiency (see Nishikawa 2019). Several students in the lower-proficiency group borrowed the words, phrases, and ideas from the source input to express their opinions. Their focus group discussions suggest that they intended to write an independent essay instead. Their misunderstanding of the task requirement could be caused by L1 transfer, specifically the norm of academic writing in Japanese. Similar findings on how L1 writing conventions may influence L2 production were also observed in other studies (e.g., Hirose 2003, Kubota 1998).

The questionnaires and focus group discussions revealed the dynamic and complex cognitive processes elicited by integrated writing. For example, the data showed how upper-intermediate level students incorporated keywords from the source texts and graphic information to summarise the main ideas in their essays. Although the questionnaire data found no statistically significant difference between the upper- and lower-proficiency groups on

the comprehensibility of visual input, the focus group discussions revealed that the less able students had faced more challenges in summarising graphic information due to a lack of lexical resources for describing the data. We argue that equipping learners with essential vocabulary knowledge for describing graphs can help improve learners' integrated writing skills.

Finally, we conclude this chapter with some implications for test developers, learners and teachers, respectively. This study only looked at Task B of the TEAP Writing test. In the TEAP test, both Tasks A and B assess integrated writing abilities. We believe Task B alone may be inadequate for assessing writing abilities at Level A2 or below. Since a summarisation task is already an integral part of Task B, adding an independent writing task as the first essay in Task B may be helpful. As shown in this study, students in high school can benefit from some scaffolding activities such as writing an outline. In addition, teachers may consider supplying learners with necessary lexical resources for describing graphic information and familiarising learners with the genres of integrated writing.

Limitations of the study

This mixed methods study illustrated the differences in cognitive processes of integrated writing between two proficiency groups with data collected from eye tracking, questionnaires, and focus group discussions. However, we acknowledge several limitations of this study. Due to a small sample size, the findings of this study may not be generalised beyond the research context. In addition, the chance of getting significant correlations could be inflated due to comparisons made between multiple pairs of variables.

It is recommended that future studies be more careful with the eye-tracking experiment design, for example, by including three or more writing tasks and taking into account the ordering effect on test performances. We also call for caution in interpreting eye-movement data without other sources of data to triangulate. It cannot be assumed that test-takers would perform better when they view certain AOIs for longer or shorter periods. For example, longer fixations may indicate difficulty in processing the information in the AOIs or efforts made to lift sentences from the source texts.

References

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

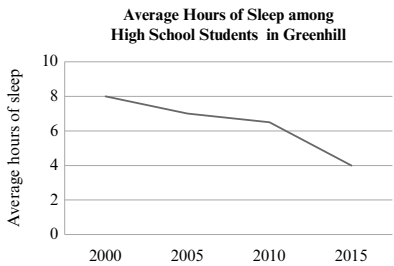
Bax, S (2013b) *Readers' cognitive processes during IELTS Reading tests: evidence from eye tracking*, IELTS Research Reports 13-06, British Council/IDP: IELTS Australia/Cambridge English Language Assessment.

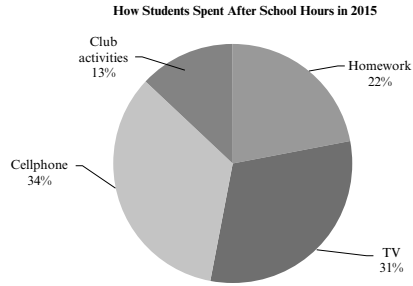
Bridges, G (2010) Demonstrating cognitive validity of IELTS Academic Writing task 1, *Research Notes* 42, 24–33.
Brunfaut, T and McCray, G (2015) *Looking into test-takers' cognitive processes whilst completing reading tasks: A mixed-method eye-tracking and stimulated recall study*, ARAGs Research Reports Online Volume AR/2015/001, London: The British Council.
Carswell, C, Emery, C and Lonon, A M (1993) Stimulus complexity and information integration in the spontaneous interpretations of line graphs, *Applied Cognitive Psychology* 7 (4), 341–357.
Chan, S (2013) *Establishing the Validity of Reading-into-writing Test Tasks for the UK Academic Context*, unpublished PhD thesis, University of Bedfordshire.
Delaney, Y A (2008) Investigating the reading-to-write construct, *Journal of English for Academic Purposes* 7 (3), 140–150.
Emig, J (1971) *The composing processes of twelfth graders*, Illinois: National Council of Teachers of English.
Emig, J (1983) *The web of meaning: Essays on writing, teaching, learning, and thinking*, New Jersey: Boynton/Cook.
Flower, L and Hayes, J R (1981) A cognitive process theory of writing, *College Composition and Communication* 32 (4), 365–387.
Grabe, W and Kaplan, R (1996) *Theory and Practice of Writing*, New York: Longman.
Hirose, K (2003) Comparing L1 and L2 organizational patterns in the argumentative writing of Japanese EFL students, *Journal of Second Language Writing* 12 (2), 181–209.
Kubota, R (1998) An investigation of Japanese and English L1 essay organization: Differences and similarities, *Canadian Modern Language Review* 54 (4), 475–508.
Nishikawa, M (2018) *Test-takers' cognitive processes while synthesizing multiple texts and graphs*, unpublished doctoral thesis, University of Bristol, UK.
Nishikawa, M (2019) Eye-tracking evidence on the role of second language proficiency in integrated writing task performance, in Papageorgiou, S and Bailey, K M (Eds) *Global Perspectives on Language Assessment: Research, Theory, and Practice*, New York: Routledge, 122–139.
Plakans, L (2008) Comparing composing processes in writing-only and reading-to-write test tasks, *Assessing Writing* 13 (2), 111–129.
Plakans, L (2009a) The role of reading strategies in integrated L2 writing tasks, *Journal of English for Academic Purposes* 8 (4), 252–266.
Plakans, L (2009b) Discourse synthesis in integrated second language writing assessment, *Language Testing* 26 (4), 561–587.
Phakiti, A (2003) A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance, *Language Testing* 20 (1), 26–56.
Scardamalia, M and Bereiter, C (1987) Knowledge telling and knowledge transforming in written composition, *Advances in Applied Psycholinguistics* 2, 142–175.
Spivey, N N (1990) Transforming Texts: Constructive Processes in Reading and Writing, *Written Communication* 7 (2), 256–287.
Suvorov, R (2015) The use of eye tracking in research on video-based second language (L2) listening assessment: A comparison of context videos and content videos, *Language Testing* 32 (4), 463–483.

Weigle, S C (2002) Assessing Writing, Cambridge: Cambridge University Press.
Weir, C J (2005) Language Testing and Validation: *An Evidence-Based Approach*, Basingstoke: Palgrave Macmillan.
Xu, Y and Wu, Z (2012) Test-taking strategies for a high-stakes writing test: An exploratory study of 12 Chinese EFL learners, *Assessing Writing* 17 (3), 174–190.
Yang, H C (2012) Modeling the relationships between test-taking strategies and test performance on a graph-writing task: Implications for EAP, *English for Specific Purposes* 31 (3), 174–187.
Yu, G (2008) Reading to summarize in English and Chinese: A tale of two languages?, *Language Testing* 25 (4), 521–551.
Yu, G (2009) The shifting sands in the effects of source text summarizability on summary writing, *Assessing Writing* 14 (2), 116–137.
Yu, G and Lin, S W (2014) *A compatibility study on the cognitive processes of taking graph based GEPT-Advanced and IELTS-Academic writing tasks*, LTTC-GEPT Research Reports RG-02, Taipei: The Language Training and Testing Center.
Yu, G, He, L and Isaacs, T (2017) *The cognitive processes of taking IELTS Academic Writing Task 1: From concurrent think aloud to eye-tracking with stimulated recall interview*, IELTS Research Reports Online Series No. 2012/2, British Council/IDP: IELTS Australia/Cambridge English Language Assessment.
Yu, G, Rea-Dickins, P and Kiely, R (2011) *The cognitive processes of taking IELTS Academic Writing Task 1*, IELTS Research Reports 11, British Council/IDP: IELTS Australia/Cambridge English Language Assessment.

Appendix 1: TEAP integrated writing used for this study (Task 1, Task 2)

Task: Your teacher has asked you to write an essay for class using the information below. Describe the situation concerning schools in Greenhill and summarize the main points about the solutions that have been suggested. In your conclusion, say which of the solutions you think would work the best based on the reason given. You should write about 200 words.





Education News

A new report found a worrying trend concerning teenagers in Greenhill. Mike Parker, the Principal at North Greenhill High School, thinks that changing the daily routine after school is the first step in dealing with the problem. Parker talked about his ideas at a recent meeting with parents. "We need to educate our children about the importance of sleep," Parker said. "Average teenagers need about nine hours of sleep each night to feel well rested," he said. Parker wants to hold a special session for students to discuss this issue.

However, Parker suggested other steps are also necessary. For example, the government is considering changing school start times in high school. One reason is that high school students tend to stay up late doing homework and other activities. "To be more realistic," said Parker, "this might be the only way to solve the problem. By delaying the start of morning classes by an hour, we can ensure that students gain extra sleep." Many teachers agree that students perform better when they are well-rested.

LETTER TO THE EDITOR

Dear Editor,

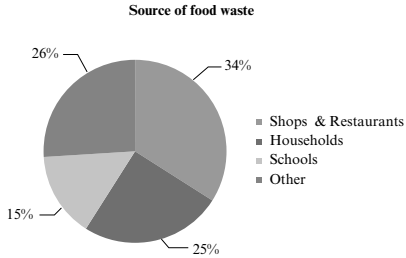
I am very concerned about the recent trend among adolescents regarding sleep. Based upon my many years of experience as a school nurse, I would like to offer some advice. First, exercise can help young people fall asleep faster and sleep more soundly. According to recent studies, it is better to work out earlier in the day than in the evening. So, it might be helpful to schedule morning exercise as a part of regular school activities in order to encourage adolescents to go to sleep before midnight.

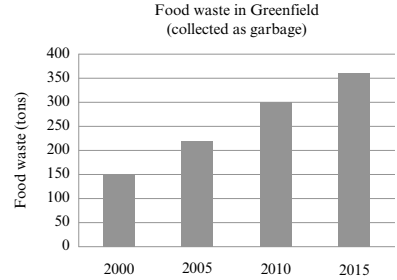
I also believe the government should take action to resolve the situation. Some experts recommend adjusting the school times to fit the biological clocks of adolescents. Research has shown that students in classes with a later start time were twice as productive as those in normal classes. It may be worth considering changing school start times in Greenhill high schools.

I'm confident that parents and local educators will find effective ways to address this issue.

Sincerely,
Sarah Case

Task: Your teacher has asked you to write an essay for class using the information below. Describe the situation concerning food waste in Greenfield and summarize the main points about the solutions that have been suggested. In your conclusion, say which of the solutions you think would work the best based on the reasons given. You should write about 200 words.





City Council News

Members of Greenfield City Council met yesterday to discuss the issue of waste food in the city. Mike Parker, the head of the city's waste collection unit expressed his concern. "The city's garbage-collection service has a very heavy workload," he said. He also pointed out that collecting and disposing of food waste has led to increased costs for the city. Parker thinks that targeting restaurants is the first step in dealing with the problem. "We need to encourage restaurants to throw away less food," he suggested, adding that "the only way to do this is to hit them where it hurts, in their pockets." To do this, he suggested introducing a special tax that would be calculated based on the amount of food waste a business produces.

Parker also put forward another idea. "The food that is thrown away by grocery stores cannot be consumed by humans," he said, "but it could be used for a different purpose. I have a friend who is a farmer, and he told me he'd welcome the unwanted food to feed his animals." Parker argued that this would be a good way to reduce costs for the council and benefit many local farmers. The city council will discuss this issue further at its next meeting.

LETTER TO THE EDITOR

Dear Editor,

I am very concerned about the recent trend in Greenfield. In my work as an environmental health officer I sometimes visit local stores. I'm always shocked by how much unsold food they throw away. I know that it is difficult for them to judge how much they will sell each day, but I believe we should try and find a way to reuse some of this food.

I also believe families in Greenfield must change their habits. When I visit my friends' homes, I'm surprised by how much food they discard without even thinking. I suggest it would be a good idea to have classes in schools to teach children about this issue. If children learn from a young age that wasting food is bad, they will grow up to teach their own children the same. This will also please parents, as their children will learn to throw away less food. I read a recent report that showed that families can reduce their monthly spending by twenty percent simply by wasting less food.

I'm confident that the city council will find effective ways to address this issue, and I hope my ideas will be useful for them.

Sincerely,
Sarah Case

Appendix 2: TEAP rating scales and criteria for integrated writing essay

Raw score (CEFR)	Main Ideas	Coherence	Cohesion	Lexical Range and Accuracy	Grammatical Range and Accuracy
+ **3** − B2	Synthesises and evaluates information and arguments from all of the verbal and nonverbal input texts.	Organised as a coherent response to the task; organisation of ideas within and across paragraphs is generally clear, though may be formulaic.	Uses discourse markers and referential cohesive devices effectively to mark the relationship between sentences and link utterances into clear, coherent discourse.	Uses appropriate synonyms and alternative expressions to convey the main ideas.	Uses a range of sentence structures appropriately; grammatical errors rarely occur and do not impede understanding of the message.
+ **2** − B1	Provides a basic summary of some of the main points, bringing together information from more than one of the input texts.	Has a logical structure but the organisation of ideas may not always be clear; organised into paragraphs, but the paragraph structure may not be completely appropriate.	Sentences and paragraphs are generally connected using discourse markers; use of referential cohesive devices (for example, pronominal reference) is mostly clear.	Gives a basic description of the main ideas in the input texts but tends to rely on the vocabulary supplied in the input texts. Some inappropriate vocabulary usage is evident.	Grammatical errors occur frequently but tend to be associated with attempts at complex structures and do not impede communication of the message.
+ **1** − A2	The response refers to some of the elements or points mentioned in one or more of the input texts (verbal and/or non-verbal) but does not synthesise these points or make clear how they are related.	No logical paragraph structure or some separation which is not appropriate; text consists of mainly unconnected sentences with no clear direction or progression across sentences.	Uses conjunctions to link clauses within sentences, but generally does not mark clearly the relationship between sentences. Use of cohesive referential devices (for example, pronominal reference) is generally not clear.	Usage of paraphrasing and synonyms is extremely limited, and alternatives are not appropriate for the task. Errors and unnatural/inappropriate usage common when reusing vocabulary from the input texts.	Grammatical errors occur systematically and may impede communication of the message.
0 Below A2	Has almost entirely copied from the input text or the number of words is too short (fewer than 50). Has written on a topic different from those assigned. Has connected to the prompt so loosely that the essay could have been prepared in advance. Requires considerable effort to see any connection being made between the composition and the prompt.				

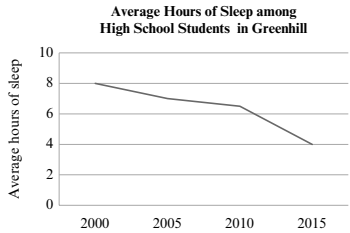
Adapted from: www.eiken.or.jp/teap/construct/rating_crit.html

Appendix 3: Questionnaire

Graph familiarity

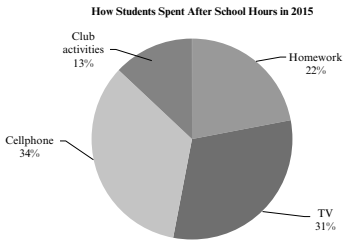
1. From the following four graphs below please rate each of the graphs in order of difficulty on a scale of 5, with # 1 being the most difficult to understand and 5 being the easiest. Please circle one that applies.

(Q2_1) Very Difficult 1 | Difficult 2 | Neutral 3 | Easy 4 | Very Easy 5

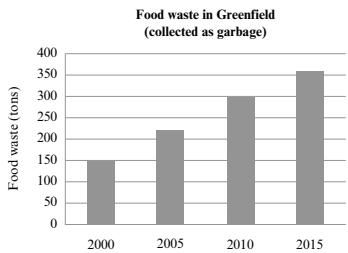


Graph familiarity

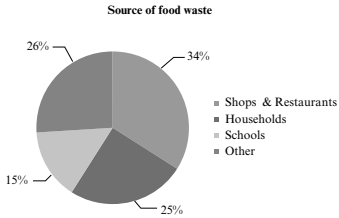
(Q2_2) Very Difficult 1 | Difficult 2 | Neutral 3 | Easy 4 | Very Easy 5



(Q2_3) Very Difficult 1 | Difficult 2 | Neutral 3 | Easy 4 | Very Easy 5



(Q2_4)	Very Difficult 1	Difficult 2	Neutral 3	Easy 4	Very Easy 5



2. How would you rate your ability to interpret the graph information in English? (Q2_5)	Very Weak 1	Weak 2	Average 3	Strong 4	Very Strong 5

Test-taking strategies

Preparing-to-Write	Related survey questions	Question number
Task Representation	Did you understand the instructions on how to write your essay?	(Q3_1)
	How easy or difficult did you find it to fulfil the task requirement?	(Q3_2)
	Which part of the task requirements did you find most challenging?	(Q3_3)
Macro-Planning [Response format]	Did you identify the purpose of the essay?	(Q4_1)
	Did you think about which solutions would work the best?	(Q4_2)
	Did you make an outline BEFORE writing your essay?	(Q7_1)
	Did you decide how many paragraphs there should be in your essay?	(Q7_2)
Micro-Planning [Graph information]	How easy or difficult was it for you to read the titles of the graphs?	(Q5_1)
	How easy or difficult was it for you to read the values on the graphs?	(Q5_2)
	How easy or difficult was it for you to read the units on the x and y axes on the graphs?	(Q5_3)
	How easy or difficult was it for you to interpret the information in the graphs?	(Q5_4)
	How easy or difficult was it for you to summarise the main trends of the graphs in English?	(Q5_5)
[Text Information]	Did you re-read the parts you thought it was important to include in your essay?	(Q6_1)
	How did you decide which information to include from the texts?	(Q6_2)
	How did you connect the ideas from the texts?	(Q6_3)

Preparing-to-Write	Related survey questions	Question number
Translation: Writing		
[Fulfilling the task requirements]	How well do you think you have described the situation?	(Q8_1)
	How well do you think you have summarised the main points?	(Q8_2)
	How well do you think you have stated which solution would work best?	(Q8_3)
	How well do you think you have given the reasons for the choice made?	(Q8_4)
[Translating: Use of language]	Did you think what verb tense form should be used before writing?	(Q8_5)
	When did you decide which verb to use?	(Q8_6)
	Which tense did you mostly use?	(Q8_7)
	Did you think which pronoun form should be used for writing?	(Q8_8)
	Which pronoun form did you mostly use?	(Q8_9)
Monitoring & Revising	I checked if my sentences were grammatically correct.	(Q9_1)
	I checked if my spelling was correct.	(Q9_2)
	I checked if I had connected the ideas from the graphs.	(Q9_3)
	I checked if I put my ideas in a logical order.	(Q9_4)
	I checked if I had fulfilled the task requirements by going back to the instruction.	(Q9_5)
	I tried my best to avoid repeating the same word or expressions in the essay.	(Q9_6)
	I checked if my essay was an appropriate length.	(Q9_7)
	I used some sentences and phrases prepared in advance to be used in the essay.	(Q9_8)

Appendix 4: Questions prepared for focus group discussion

The following questions were discussed in all groups.

a. Did you clearly understand the instruction of the test?
b. Which part of the tasks did you have the most difficulty with?
   - Reading the texts
   - Understanding the graphs
   - Describing the situation by interpreting the graph information
   - Summarising the main points about solutions that have been suggested
   - Choosing the best solution
   - Giving the reasons behind your argument
c. To what extent do you think your typing skills of the keyboard affected your writing process?
d. How did you decide which information to include?
e. How did you plan when to start to write your essay?
f. What did you often do while you were writing?
g. What did you do after you finished writing your essay?
h. What kind of strategies did you use while taking the test?
i. What were you mostly concerned about with your essay?
j. What do you think you would do differently if you had to take the same test again?

7 Comparing methods to identify test-takers' attention to diagnostic feedback on an English reading test

Maggie Dunlop
Ontario Institute for Studies in Education, University of Toronto, Canada

Abstract

Eye tracking has a strong history in reading research and is gaining popularity in the field of language assessment. It is used to identify attention, or at least visual attending, to visual stimuli. As eye tracking grows as a means of validating language assessments, the discussion about what 'attention' means, and how to measure it, has matured in the language assessment community. This chapter reviews some different ways the language assessment community measures attention, for example eye tracking and stimulated recall interviews. The chapter then looks at the findings of a study in which three methods were used to measure attention of adult immigrant English language learners to written diagnostic feedback on an English language reading test: eye tracking immediately followed by think-aloud interviews, self-reports on a questionnaire, and one-month-delayed recall interviews. The chapter investigates how different data collection methods access different aspects of the construct of 'attention' for processing written language learning feedback, and how these differences might inform the way eye-tracking data is used in validation activities for language assessments designed to support learning.

As the three methods yielded three very different sets of results about participants' 'attention' to feedback, the chapter discusses how these results all yield useful, but different, information regarding such attention. It concludes by discussing the ways that the choice of research question about 'attention' to feedback will therefore influence data collection method (or vice versa), and the implications for how eye tracking and other measures of 'attention' usefully and less usefully contribute to validation activities that involve written diagnostic feedback about second language skills.

Introduction

This chapter starts from the premise that relevant feedback on second language skills is a crucial part of second language learning. In fact, for any assessment that claims to inform ongoing learning in some way (rather than simply fulfilling certification requirements), it might be argued that omitting relevant feedback nullifies any validity claims about other features the assessment may support. This is because if learners cannot derive any usable information from the assessment, they may as well have not done the assessment. And indeed, in terms of validity frameworks, the above proposition is increasingly acknowledged, and some work is emerging in which the quality of feedback is the central validity claim (Chapelle, Cotos and Lee 2015, Jones and Saville 2016).

As a result of growing consensus around the critical role of feedback in many language testing contexts, research on processing and usage of feedback in language assessments is gaining ground (Fernández-Toro and Hurd 2014, Jang, Dunlop, Park and van der Boom 2015, Wagner 2015). That said, practitioners still lack evidence on which to base design decisions when developing feedback systems for language learning. And in turn, feedback researchers lack information about which methods yield the most robust information about various aspects of attention and cognition when processing feedback from language assessments, despite recognition that a construct can be influenced by the data collection method (Bachman 2007). Of particular interest is eye tracking, a method of accessing attention long used in reading and marketing research (e.g., Just and Carpenter 1976, 1980), and now being taken up in language testing (e.g., Bax 2013, Brunfaut and McCray 2015, Owen 2016).

This chapter addresses this methodological gap by detailing a study that was part of a larger study into use of feedback derived from performance on an English reading proficiency test. This smaller study investigated how using different methods to collect different types of data about attention yielded insight into what aspects of 'attention' each method accessed. Methods included eye tracking immediately followed by think-aloud interviews, self-reports on a questionnaire, and one-month-delayed recall interviews.

The methods yielded three very different sets of results about participants' 'attention' to feedback, and this chapter discusses how these results make sense given the means of measurement. The chapter concludes by discussing how the choice of research question about 'attention' to feedback from a language test will influence data collection method (or vice versa, Bachman 2007). The chapter also discusses ways to define the construct of 'attention' to language assessment feedback as measured by eye tracking, and the ways that eye tracking usefully and less usefully contributes to validation activities involving reporting test-takers' attention to diagnostic feedback about their language skills.

Literature review

Accessing attention

When receiving feedback, it is essential that learners process the information provided, and in order to process it, they need to pay attention, although not necessarily consciously (Duchowski 2007). Where learners fail to process information, even where the information is good quality, feedback cannot be integrated into their knowledge frameworks, and therefore cannot be retained and applied to progress learning. Two ways to access learners' attention is by tracking their gaze, and by asking them to report it.

Just and Carpenter (1976, 1980) were early leaders in the relationship between cognitive processing, attention, and visual gaze. They noted that longer fixations occurred when there were greater processing loads, and that differences in working memory might result in differing abilities to process cognitive load. In time, fixations were identified as one of the key physical measures of gaze (Duchowski 2007). Research in the 1980s also yielded that while attention and visual gaze are highly related, attention can focus on areas outside the direct gaze, or outside the gaze altogether (Posner 1980, Rayner 1998, Richardson, Dale and Spivey 2007).

As a result, although the precise physical mechanisms of visual gaze remain under some debate (Duchowski 2007), eye tracking has become an established method for exploring attention. The first work was done with reading in a first language (Just and Carpenter 1976, 1980), but the field of second language learning has taken up use of the emerging technology (Bisson, van Heuven, Conklin and Tunney 2014, Kang 2014, Winke, Sydorenko and Gass 2013). For example, Kang (2014) investigated reading strategies among first and second language readers and found that the second language readers read much more slowly, but the two groups were otherwise alike in attention distribution and reading comprehension, indicating that both groups may be utilising similar cognitive processes.

Eye tracking in language testing

Recently, language testing researchers have begun to explore via eye tracking how language test-takers engage with test tasks. For example, Bax (2013) found that eye tracking yielded valuable information about how stronger and weaker adult test-takers interacted with some second language reading test items, while Ballard and Lee (2015) made the same findings with young learners.

Suvorov (2015) also used eye tracking, in this case to understand video usage in second language listening test tasks. Language learners used the visual information to aid comprehension, and usage of videos differed

according to how the videos were structured. However, the authors did not locate research on uses of eye tracking to explore attention to feedback or processing of feedback from second or foreign language tests.

Stimulated recall interviews in language learning

In addition, stimulated recall interviews have proven effective at eliciting evidence of a range of cognitive processes in second language learning. For example, Bao, Egi and Han (2011) studied learner noticing of recasts (where a fluent speaker correctly rephrases a learner's spoken error as part of natural conversation), and found that stimulated recall interviews with learners yielded higher rates of noticing than testing learners' ability to correctly reproduce the recasts. Egi (2008) found that stimulated recall interviews used in second language acquisition research did not appear to unduly impact participants' performance on post-test measures. Zhao (2010) investigated usage and understanding of peer and teacher feedback by English as a foreign language (EFL) learners in China, and found that the interviews yielded useful information on learners' understanding of feedback. Lam (2008) even advocated using stimulated recall in English as a second language (ESL) classrooms to provide a window on learners' metacognition.

Gass and Mackey (2016:132) provide a detailed discussion of the methodological strengths and weaknesses of stimulated recall interviews for language learning research. They conclude that 'stimulated recall data can provide valuable information about some of the complex processes involved in learning L2s'. However, they also note that stimulated recall data 'must always be interpreted within the framework of current theoretical concerns, and in conjunction with other compatible and reliable data'. Indeed, it should be noted that interview methods investigating cognition are limited by needing to infer processes based on reported thoughts, and processes that are fully automatised may not emerge at all in reported thoughts (Pintrich 2004).

Combining eye tracking and stimulated recall interviews

A potential source of 'other compatible and reliable data' is eye tracking. In contrast to studies using stimulated recall methods, eye-tracking methodologies often require no verbal contributions from participants. However, a number of studies have combined the use of eye-tracking traces with verbal reports of cognition from participants (Brunfaut and McCray 2015, Godfroid and Spino 2015, Holmqvist et al 2011). Sometimes these combined approaches involve using an eye tracker while the participant verbally reports their actions and cognition (cf. Smith 2012); in other studies (cf. Brunfaut and McCray 2015, Godfroid and Schmidtke 2013) the two

methods are separated, with the verbalisations taking place after eye tracking has finished. Another methodological variation concerns the use of eye-tracking traces to prompt participant recollections. Bax (2013) used traces to prompt recollections, whereas Owen (2016) and Brunfaut and McCray (2015) note the benefits of first, prompting without the use of traces, and then prompting again with the use of traces. In general, the relative benefits for each methodological choice probably depend on the research question at hand.

In summary, existing work regarding eliciting of cognition in language assessment has utilised various methods, including eye tracking and interviewing. However, this work has not yet explicitly evaluated how these methods access different attentional processes when learners are processing feedback from a language test. A question that merits further exploration is: How might different data collection methods access different aspects of the construct of 'attention' for processing written language learning feedback, and how might this difference impact the way eye-tracking data is used in validation processes for language assessments that inform learning?

It is this question that is addressed in the study described below.

Methodology

Participants

People were invited to participate in the study if they had been in Canada for less than one year and intended to live in Canada long term, and considered themselves non-native speakers of English and were still learning English. People believed to have intermediate levels of English language proficiency, in this case Canadian Language Benchmark (CLB, Centre for Canadian Language Benchmarks 2012) 5 to 8, were invited to participate in the study, although some people with CLB 4 accreditation were also accepted into the study upon request.

Participants volunteered via invitations received either at government-funded adult ESL programmes or through word-of-mouth networks. In total, 102 people participated in the basic study, with 15 people also receiving their report while using an eye tracker then participating in a stimulated recall interview, and 15 people also participating in a delayed recall interview one month later. (The interviews overlapped but were not exactly the same; five people completed both interviews.)

The participants came from five ESL programmes (n = 80) and three word-of-mouth networks (n = 22). Almost three-quarters (74%) of participants were female and the median age was 38, with participant ages ranging from 17 to 69. As would be expected given Canada's immigration policies, which prioritise individuals with high levels of education, a little over two-thirds of

the participants (69%) held an undergraduate degree and a little over one-third (36%) also held a graduate degree. Most participants (79%) had not completed any of their education in Canada.

Reflecting general Canadian immigration trends, which includes immigration from all regions of the world, the 102 participants reported knowledge of 31 languages (including English) and reported 24 first languages. The largest first language groups were speakers of Chinese languages (39%), followed by Farsi (14%) and Spanish (13%) speakers. Languages of the Indian subcontinent represented 9% of first language speakers.

Data collection procedures

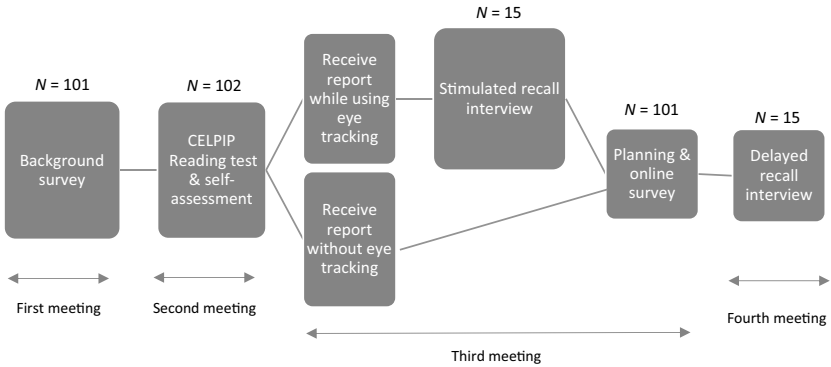
Data collection took place between January and October 2015. An outline of the study's data collection activities is presented in Figure 1.

Data collection was initiated and completed within two weeks. At the first meeting, people who consented to participate in the study completed a background survey that asked them about their demographics and English language use, and elicited information about two types of psychological characteristic: goal orientations and beliefs about intelligence.

At the second meeting, participants did the Canadian English Language Proficiency Index Program (CELPIP) Reading test under exam conditions, and then immediately after completing the test also completed a self-assessment based on their perceptions of their performance on the test. (See the section entitled 'Instruments' for more details on the CELPIP and self-assessment.)

A day or so later, most participants received their feedback reports via an online platform, completed the report planning section, completed a survey that aimed to explore their interactions with the report, and downloaded the

Figure 1 Data collection activities and order*



**CELPIP = Canadian English Language Proficiency Index Program*

report to keep. The exceptions were 15 participants who met individually with the researcher and received their reports using a computer that was integrated with eye-tracking technology. These participants then took part in stimulated recall interviews immediately after eye tracking took place. After completing the interview, these participants (along with the rest of the participants) then completed the post-report survey online and downloaded their report.

For most participants, this was the conclusion of their participation in the study. However, 15 participants also met individually one further time, a month after receiving their reports. These participants took part in a delayed recall interview. Without prior reminding of their report content, participants were asked to recall details about their report and prompted to elaborate.

Instruments

CELPIP Reading test

The CELPIP is a computer-based test of general English language proficiency developed by Paragon Testing Enterprises. It is aligned to the CLB. At the time of the study, it was one of two tests accepted by the Canadian federal government for immigration purposes in Canada. The reading test form used in this study was a retired test form from a previously operational CELPIP test. The reading test consisted of four tasks that required reading several text types and responding to 38 multiple-choice items that were binary scored correct/incorrect. Each task was timed.

Self-assessment

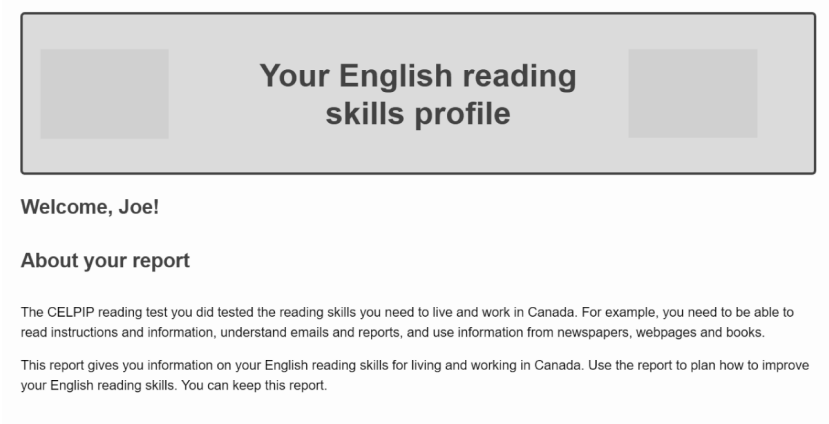
The self-assessment used 18 descriptors that were developed for the feedback report to describe the six reading skills reported – three for each skill (see Figure 3). The report descriptors were developed in collaboration with ESL teachers. On the self-assessment, the descriptors were randomised and participants were asked to rate on a 5-point Likert-style scale how often they could do each activity while doing the test. The self-assessment was paper based.

Feedback report

Each participant received an online report on one page and learners used the scroll function to read through. The report was designed to provide substantive, useful information for learning English, specifically for improving English reading skills. In order to meet this goal, the report consisted of several components. First, an introduction (see Figure 2) provided an orientation for the learner.

Learners were then presented with their test and self-assessment results. (Note that due to space limitations, how results were derived is not discussed

Figure 2 Example of introduction to report (fictional learner)



in this chapter.) Next, learners were presented with test and self-assessment results for six skills: Using vocabulary, Using directly stated information, Using indirectly stated information, Making connections, Separating ideas, and Using culture. For each skill reported, the report gave a plain-language title for the skill, three bullet points listing 'can do' statements that described what the skill was, and a figure that showed the probability of mastery alongside the learner's self-diagnosed probability of mastery. Figure 3 shows an example report of a fictional learner.

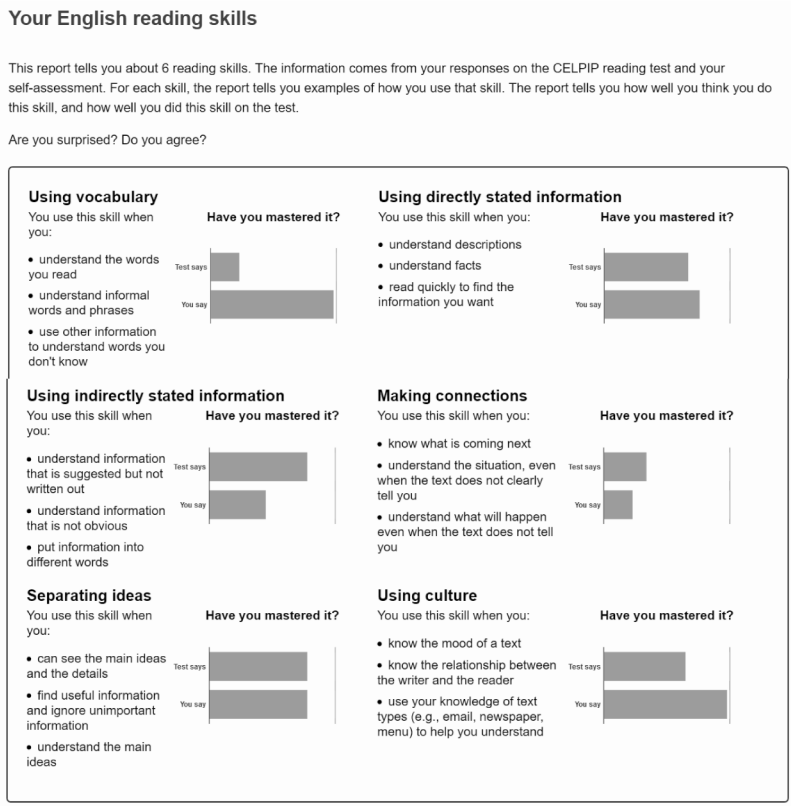
Following the skill descriptions and figures, the report provided some learning suggestions (see Figure 4). Each learner was suggested to work on two skills, and the suggestions varied according to the test-based skill profile of the individual. For each skill, three activities were suggested that would help the learner improve that skill.

Finally, participants were required to engage in some planning for their learning. They did this on their own as part of reading the report. Learners were required to select between one and three of the skills that were presented in the report, and identify learning goals for that skill, action plans for how they would achieve their goals, and monitoring plans so they would know how they were making progress. The planning section introduction and the first plan outline are shown in Figure 5; the next two plan outlines are the same.

Eye tracking

The study used a Tobii 2150 with a 21-inch monitor eye-tracking device, Tobii Studio software and a 50 hertz (Hz) sampling rate. The report was presented across six frames (see Figures 3 to 5). An eye-tracking protocol introduced the device to participants, calibrated the eye tracker, delivered the

Figure 3 Example of report skill descriptions and learner personal results (fictional learner)



report, and set guidelines for the stimulated recall interview. The participants first viewed their report silently, with the eye tracker collecting data. Next, two rounds of stimulated recall interviews took place. In the first round, the participant recalled thought processes while viewing only their report. In the second round, the participants recalled thought processes while viewing eye-tracking traces superimposed on their report, to further prompt recall. Altogether, the eye tracking and stimulated recall interview took about one hour to complete.

Fixations recorded for each participant during eye tracking were grouped according to which section of the report they fell in: the introduction, skill descriptions, figures, suggestions, or planning. Fixations were analysed descriptively (see Table 1 for summary statistics). There was much variation among participants regarding how much time they had spent looking at the report, so the percentage of time spent looking at each section was calculated,

Figure 4 Example of report suggestions section

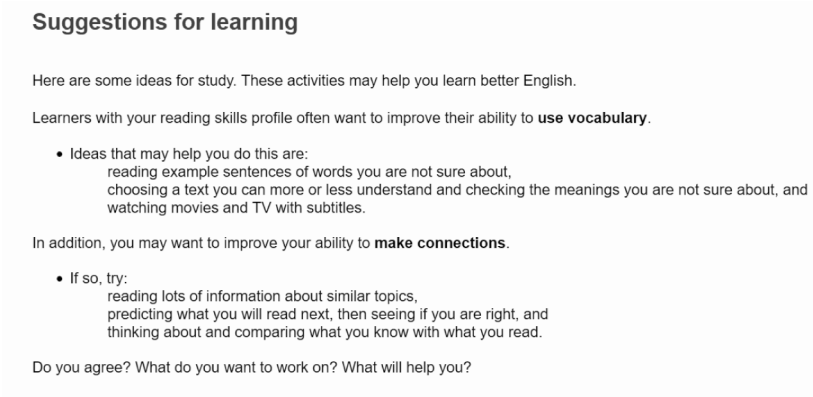
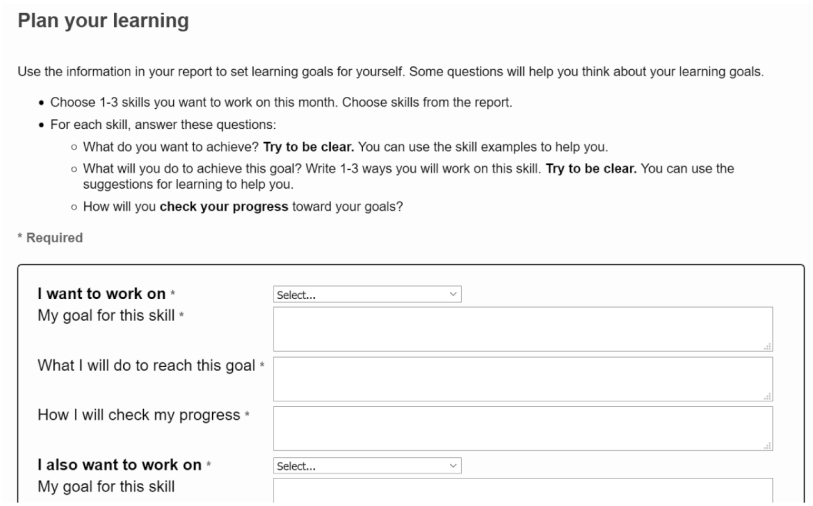


Figure 5 Example of report planning section (first plan outline only)



in order to identify which sections received relatively more attention. Eye-tracking paths were also viewed in order to identify similarities and variations in how participants read through their reports.

Post-report online survey

In a post-report online survey, the construct of reported attention was addressed by asking participants how much they looked at each section of the report: introduction, skill descriptions, figures, suggestions, and planning.

Table 1 Mean total fixation times and frequencies for each section

Report section	Mean total fixation time in seconds (SD)*	Mean count of fixations (SD)
Introduction	24.6 (10.28)	70.7 (16.72)
Descriptions	71.0 (35.72)	185.3 (80.54)
Figures	10.9 (5.13)	31.0 (11.67)
Suggestions	60.5 (38.97)	182.5 (100.56)
Planning	62.2 (25.44)	187.3 (65.02)

**SD = Standard deviation*

Participants responded on a Likert-style 1 to 5 scale of 'I did not look at this', 'A little time', 'Some time', 'A lot of time' and 'All my time', or could select 'What is this?'.

Delayed recall interview

The delayed recall interview aimed to find out which aspects of the report were sufficiently paid attention to that they were recalled without prompting one month later. The interview took place one month after participants received their report. A protocol structured the recall interview to be a semi-structured interview, with six questions to prompt recall of report usage and content:

1. How have you used your reading report?
2. What do you remember about your reading report?
3. What did the report tell you about your reading skills?
4. What did the report look like?
5. What did your report suggest you should do to continue learning?
6. What learning goals did you set?

Interview transcripts were coded for any references to report sections.

Results

Self-reported attention to the report

The first data set was formed from participants' self-reported attention to different areas of their reports. Immediately after receiving their reports and completing the planning section, all participants were asked on a survey how much they looked at each section of the report: the introduction, the skill descriptions, the figures comparing test and self-assessment scores, the learning suggestions, and the planning opportunity.

They were invited to respond on a 1 to 5 scale of 'I did not look at this', 'a little time', 'some time', 'a lot of time', and 'all my time'. They could also select 'What is this?'.

The suggestions for learning were reported to be looked at the most, with 58% of participants reporting that they spent 'a lot of time' or 'all my time' on this section. The suggestions were also the only section in which everyone reported spending at least some time. The planning section also received a lot of attention, with 52% of participants reporting that they spent 'a lot of time' or 'all my time' on this section.

The section receiving third-most perceived attention was the figures that compared test and self-assessment scores. In total, 37% of participants reported spending 'a lot of time' or 'all my time' on this section. The introduction was perceived to receive a similar amount of attention; 31% of participants reported spending 'a lot of time' or 'all my time' on this section. In contrast, just 25% of participants reported spending 'a lot of time' or 'all my time' on the skill descriptions.

In addition to the survey, 12 participants also completed interviews in which they were shown eye-tracking traces of their report reading and asked to recall their thoughts. The depth and extent of cognition reported by participants consistently aligned strongly with the amount of time they reported looking at each section of the report.

A typical example of this relationship is shown using participant DP's responses. DP reported spending 'a little time' looking at the introduction and 'some time' looking at the skill descriptions, but 'a lot of time' looking at the figures, suggestions and planning. Furthermore, DP's reported cognition for the introduction was represented in statements such as: 'No I didn't, I just look through…'. Likewise, DP rarely commented on skill descriptions, and when he did his comments were minimal, for example (comments are unedited to maintain authenticity): 'I just read and read … I usually don't, like, pay too much attention to the word, yeah.' In contrast, DP reported substantial cognition around the figures, for example: 'The test said, because I want to see what the test says, yeah.' DP also frequently engaged with the suggestions and planning, for example noting the suggestions' value: 'I think maybe it's the most important part for me, and I can get some good information and helpful information from this page. So I take this page seriously…'.

Actual time spent looking at the report

The second data set is formed from eye-tracking traces for 11 participants who completed eye tracking/stimulated recall interviews and whose data was usable. (Four other participants completed the interview but either the eye-tracking data collected was of insufficient quality, or participants' low

English language proficiency precluded effective thinking aloud in English.) Duration and number of fixations in each area indicated which areas of the report participants looked at most.

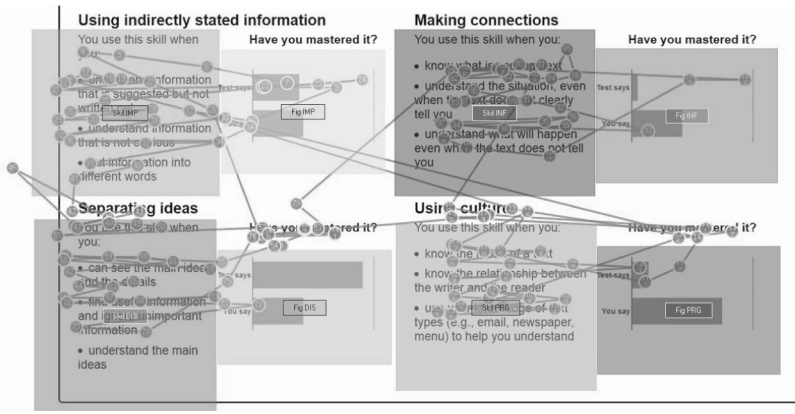
The fixation data contrasted with the reported time spent looking at report sections. Participants reported spending a lot of time looking at suggestions and planning – ranking first and second in amount of reported time, but skill descriptions placed last (fifth). However, participants' eye-tracking traces indicate that on average, participants spent about 25% of their time on suggestions, 25% on planning, and slightly more on skill descriptions. The fixation data indicated participants in fact spent a lot of time looking at the skill descriptions.

Similarly, the fixation data indicated that on average participants spent only 5% of their time looking at figures, a clear bottom (fifth) rank for amount of observed attention. Figure 6 shows an example of how fixations contrasted between skill descriptions and figures. However, this observation is in direct contrast to the amount of time reported for this section; 75% of participants reported spending 'some' to 'all my time' on this section and it ranked third in amount of reported attention.

The only report section in which observed time resembled reported time was the introduction. The amount of time spent on the introduction was on average 10% of total time, which placed it in fourth rank for amount of attention, comparable to the reported amount of time, in which it also placed joint-fourth rank.

Overall, these findings indicate that the amount of time spent on a section was most strongly related with the amount of text in the section. The skill descriptions, followed by the suggestions and planning, were the most text-heavy section. The figures had very little text.

Figure 6 Eye-tracking traces for a participant with mixed skill mastery



Notably, other information was derived from the eye-tracking traces that was not directly comparable to other data sets. The eye-tracking traces provided information about the order in which participants viewed report sections, and how often they visited them. This information yielded findings such as the relative utility of figures. For example, skill profiles that included a mix of higher and lower skill masteries incited the most frequent reference to the figures whereas consistently low skill profiles incited ignoring of the figures. The skill mastery bar appeared to be used as a proxy for traditional 'test results'.

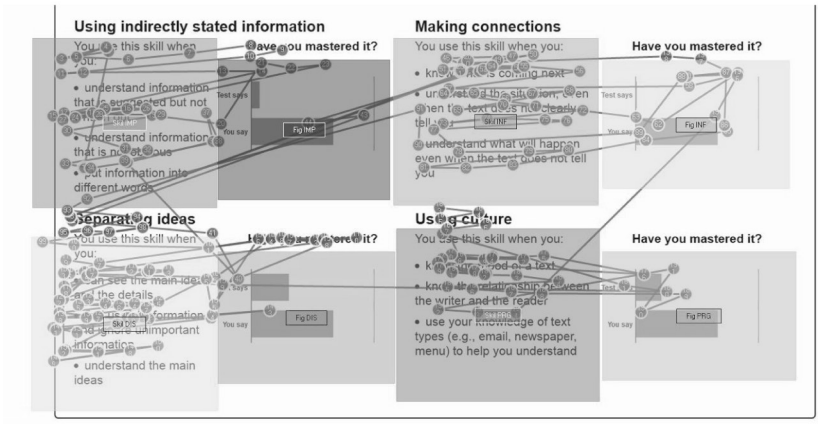
Another finding was how participants used page layout. Figure 7 presents an example of the differences in fixations for skill titles and descriptions. The eye-tracking traces indicated that participants only looked at skill titles as part of the flow of text, and did not use the skill titles either to orient themselves or make meaning of the bullet points below. In contrast, there was consistent interest in looking at the bullet points, possibly to make meaning of the figures.

Report features that could be recalled after one month

The third data set is formed from the responses of 15 participants who completed delayed recall interviews. One month after receiving their report, these participants recalled what they remembered of the report and discussed how they had been using it.

A hierarchy emerged regarding which areas of the report participants recalled. The figures were clearly the most remembered part of the report.

Figure 7 Eye-tracking traces showing uneven fixation distribution on skill titles



Of the 15 participants, 12 recalled the figures (i.e., results). The three who failed to mention the figures or results recalled very little at all from the report. Comments often recalled discrepancies between skills or between the self-assessment and test. For example:

> I remember there is a, I think, graphic that said I'm weak in this or I get for example less maybe in reading. And it was some marks good and some marks weak.

Others interpreted the differences as low achievement, for example:

> Some report tell me the reading is I think more poor because I understand a bit a little something.

The next most-recalled areas of the report were the planning section and the suggestions for learning. Seven participants recalled one or both sections, and all seven also recalled the figures. The suggestions recalled were ones that participants found personally meaningful, for example:

> The report told me how can I improve my reading skills. Read the newspaper or something else, and listening to radio and watching TV…

> Read some label, about label something, label, so sometime I will, when I went to the store I will read the label.

More detailed suggestions regarding habits while reading, for example predicting what will come next, were not recalled by any participants.

The specificity with which plans were recalled varied, and was related to the specificity of the original plans. For example, the participant who originally wrote:

> I don't understand many difficult vocabulary. So I must study it. Every day I will read a newspaper and check difficult words

recalled:

> Yeah so I want to know about the vocabulary…

In contrast, the participant who originally wrote:

> I will help my children to study at schools with their homework; I will continue taking my LINC classes; I will widen my communication with native speakers; I will continue any kind of reading, listening, watching and using.

recalled:

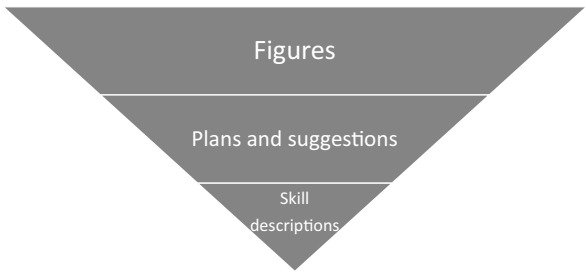
> I say should continue taking our classes here, LINC classes, and maybe listening news, talking in the communities, maybe finding networking to have practice for conversation. I believe that I wrote that doing homework with my kids helps me too, so maybe I wrote something like that.

Finally, three participants recalled something about the skill descriptions, mainly the names of the skills, although no one recalled all six skill names. One participant recalled some of the bullet points in the descriptions. All three of these participants also recalled the figures, suggestions and plans, and reported using the report in some detail. It appears that recall of the skill descriptions is at the bottom of an inverted hierarchy of recall, with engagement in other areas of the report required before skill descriptions are processed sufficiently to be recalled one month later.

The inverted hierarchy of recall for the report sections was therefore figures, plans and suggestions, and skill descriptions, shown in Figure 8. None of the participants mentioned the introduction. Recall of the figures or the information contained in the figures appeared to be a prerequisite for being able to recall anything from the report. Plans and suggestions were recalled in different ways, and finally descriptions could be recalled if there was enough investment in the content of the report.

This inverted hierarchy indicates the extent personalisation affected long-term recall. The figures contained highly personal information, including participants' test results, their self-assessment results, and a comparison of the discrepancies between them. The plans, which were made by the participants themselves, were also personal and additionally were self-determined, likely facilitating long-term recall. The suggestions deemed by participants to be relevant to their lives were also likely to be recalled. In contrast, the skill descriptions were only recalled to the extent that they provided meaning to participants' understanding of the report, and were

Figure 8 Hierarchy of recall for report sections



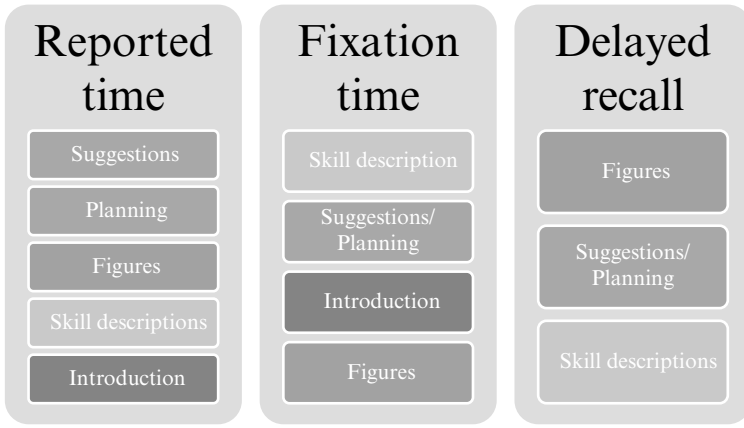
only recalled by participants who had clearly devoted substantial attention to the report. Finally, the introduction, which was generic (with the exception of the participants' name), was not mentioned at all.

Discussion

Before moving on to discuss results, it is useful to acknowledge the limitations of the study design and thus the results too. Participants who participated in eye tracking received their report split into six screens, and could not immediately write their learning plans, so the eye-tracking and stimulated recall interview data could not exactly reflect the processes of participants receiving their report online. This group of participants also completed their survey after finishing eye tracking and interview activities, which likely impacted their survey responses. Participants in delayed recall interviews one month later had all completed the survey and some had participated in eye tracking and a stimulated recall interview, which may have influenced earlier processing and therefore potentially also their responses in the delayed recall interviews. Finally, due to logistical constraints the report and study instruments were all in English, the target study language. Using English clearly affected processing of report content, especially among participants with lower English proficiency, and likely also impacted participants' articulation of thoughts, and their ability to understand the survey.

Within these limitations, the findings described in the results indicate that reported and observed attention to written feedback for a language test diverged substantially, and this divergence is summarised in Figure 9.

Figure 9 Relative amount of attention noted in the relevant data sources: Sections with most attention listed first



When asked which areas of the feedback report they spent the most time on, participants reported spending the most time on suggestions and planning (M = 3.61 and 3.42, SD = 0.93 and 1.07 respectively, on a scale of 1 to 5), followed by figures (M = 3.20, SD = 0.99). The observed time spent on each report section contradicted these claims, as the most time was clearly spent on skill descriptions (28% of time on average), closely followed by suggestions and planning (25% and 24% of time on average, respectively). In the observed data, it was the figures that participants spent very little time looking at (5% of time on average).

This contradiction can be explained from the eye-tracking/stimulated recall interview data. The data indicated that the amount of time reported on each report section was closely related to the amount and depth of cognition undertaken by participants for that section. For example, as intermediate-level English language learners, participants needed to focus carefully on comprehending the skill descriptions. Therefore, it is reasonable to conclude that when asked to report time spent, participants reported the amount of thinking that they recalled doing for each section, which was not the same as the amount of time they spent on each section.

However, the third data set concerning attention – the delayed recall interviews – added further information about the figures. Although participants reported a moderate amount of attention to the figures and were observed to spend very little time looking at the figures, the figures and the information in them were clearly the most memorable information on the report, recalled by 12 out of 15 delayed recall interview participants (80%). In contrast, suggestions for learning and planning were recalled by a similar proportion (50% and 57% respectively) of participants as those who reported spending 'a lot of time' or 'all my time' on these sections (58% and 52% respectively). Likewise, only three participants (27%) recalled skill descriptions, and a similar percentage (25%) reported spending 'a lot of time' or 'all my time' on these sections. Note also that the eye-tracking trace data explains this discrepancy for figures by showing how participants viewed the figures. Although little time was spent looking at them, they were a 'framework' for interpreting other information in the report, with users switching back and forth – if the figures had useful information – between figures and other available information. Possibly, participants were interpreting and using figures in similar ways that they would interpret and use test scores.

Therefore, although participant groups for each data source did not completely overlap and there is some confounding among the three data sources due to study design limitations, the relationships between reported time spent looking at each section, eye tracking/stimulated recall interviews, and the delayed recall interviews appear to indicate that adult language learners can report attention to feedback reports on a language proficiency

test fairly reliably. Note that in contrast, the amount of time spent on a section is not directly related to attention; text-heavy sections simply require more time to read. In fact, there was an inverse relationship between amount of text in a section and the chances of that section being recalled one month later. However, the amount of text was also very much related to extent of personalisation; the unpersonalised sections were very unlikely to be recalled. Moreover, there was a skill profile effect; lower English reading proficiency participants had flat skill profiles and therefore paid little attention to the figures due to absence of useful information.

In conclusion, it appears that participants paid most attention to information on the feedback report that helped their language learning (suggestions and planning) at the time of receiving the report and did indeed often recall these sections subsequently. However, the section with the most powerful long-term recall was the figures, which had been used as a frame of reference when interpreting other information in the report such as the suggestions for learning and planning, and thus left a strong impression. In terms of observed and reported attention, this study's findings cohere with existing knowledge of visual processing and attention (Bisson et al 2014, Kang 2014, Owen 2016, Wadlinger and Isaacowitz 2011), and have implications for the design of language learning feedback and the usage of various stimulated recall methods.

Regarding why participants did not accurately report the actual time that they spent looking at each report section, research points to the automaticity of many decisions about directing attention (Uusberg, Uibo, Kreegipuu and Tamm 2013). First, the eye-tracking data clearly reflects levels of cognitive demand and processing weight, with more linguistically demanding report sections taking more time to read. This finding confirms existing research in cognitive load (cf. Ayres and Paas 2012), which makes the claim that the ability to process information is affected by the relative load that the individual experiences at that time (Sweller, Ayres and Kalyuga 2011). A primary source of cognitive load is unfamiliar text conveying abstract information, particularly when compared to figures, and particularly for language learners (Fontanini and Braga Tomitch 2009, Segalowitz and Frenkiel-Fishman 2005). Indeed, Kang's (2014) work in reading strategies likewise notes that second language learners read more slowly than first language readers.

However, the self-reported time spent on each area of the feedback report shortly after receiving the report appears to primarily reflect participants' actual attention rather than actual time spent looking at each section. At this point the well-known differences between attention and gaze (Posner 1980, Rayner 1998, Richardson et al 2007) become apparent, and the automaticity effect comes into play (Uusberg et al 2013). The principal aspect to note is the relationship of affect with attention and processing,

known as affective attention (Uusberg et al 2013). Essentially, as Kissler, Herbert, Winkler and Junghofer (2009:75) note, 'stimuli that people regard as emotionally arousing obtain prioritised processing'. This phenomenon is believed to be a survival mechanism to ensure appropriate response to both negative and positive stimuli. Within this theory, personalised information about one's language abilities and learning would be likely to receive more attention than general information, which was observed in this study. Combining cognitive load theory and affective attention theory suggests why suggestions and planning were self-reported for more time/attention than the figures. To an extent, individuals are conscious of the time and effort they devote to something because they can sense greater and lesser cognitive load.

Finally, building on affective attention theory, as external judgements of one's competency hit hard at one's personal ego and self-efficacy (Ertac 2011), it can be argued that figures depicting language learning achievement to an immigrant language learner, where the stakes for success are high, are likely to be particularly arousing and therefore impactful. This phenomenon was observed in this study through the prevalent long-term recall of figures. The findings also reflect Smither, Brett and Atwater's (2008) work outside language assessment on employees' recall of feedback wherein employees were likely to recall supervisor or external feedback – once again, external, authoritative judgements made a substantial long-term impact on individuals. That the personalised suggestions for learning and planning opportunities were secondarily also recalled, and non-personalised information was not recalled at all, indicates the depth of impact that personalised assessment feedback can have, as noted above.

Conclusion

Even given the methodological limitations of the study, the discrepancies that the study found between data collection methods do not appear to originate from reporting inaccuracies by either participants or the eye tracker, but that each measure was tapping into different aspects of attention and processing. These findings inform how useful eye tracking and other methods are for understanding 'attention' to written language testing feedback.

Many empirical measures, such as eye tracking, measures of biodata and even facial recognition, provide objective data. However, this study's results confirm existing research in other areas which has found that processes occurring in the mind are often beyond the reach of these measures. Similarly, the study found that subjective measures designed to access unevaluated cognition during processing of feedback such as think-aloud and recall interviews, as well as surveys and interviews designed to collect self-report

evaluations, yielded data that well reflected (and better reflected) relevant processes within the mind. This finding also coheres with research in other areas.

A key implication for researchers wishing to understand 'attention' to written language testing feedback is that careful consideration of the definition of 'attention' is required, as is selecting (an) appropriate data collection method(s). Eye tracking appears most useful for understanding attention in terms of 'attending to the text' in a feedback report. This type of information about attention facilitates understanding cognitive processing load (e.g., amount/complexity of text), and exploring how readers utilise different structural features of a page to build meaning and scaffold interpretation (e.g., titles, figures, descriptive text). Such information is very important for designing feedback that is easily comprehended, digested, and perceived as useful, characteristics likely to be valued within validation approaches focusing on feedback quality.

However, if 'mind's eye' attention to a report is the key interest, eye tracking comes up short. In contrast, traditional methods such as (well-designed) surveys and think-aloud, recall and structured interviews appear to deliver robust data regarding what impacts people emotionally/attentionally when they read a feedback report based on a recent language test. That is, what they regard as significant and meaningful, and what makes the most emotional impact. Validation activities that require understanding of people's emotional experiences of and thinking strategies for diagnostic feedback about language skills may want to focus on traditional methods.

Finally, the study found the time of data collection is important; people report information that is important to them at that time. Therefore, researchers of language learning feedback need to consider exactly what aspects of a validation process for language assessment feedback their research may inform. For example, is the researcher aiming to evaluate how well learners digest language learning feedback upon receipt of it, or to evaluate learners' longer-term usage of feedback? The results of this study indicate that different methods are appropriate to find answers to these different questions.

Acknowledgements

This research was possible due to financial support from OISE/University of Toronto Funding Grants and Academic Excellence Awards, Ontario Graduate Scholarships, a CELPIP-General Doctoral Studies Research Grant, a Paragon Research Award, and a TOEFL Small Grant for Doctoral Research in Second or Foreign Language Assessment. Logistical support from Paragon Testing Enterprises was also key to completing the research.

References

Ayres, P and Paas, F (2012) Cognitive load theory: New directions and challenges, *Applied Cognitive Psychology* 26 (6), 827–832.

Bachman, L F (2007) What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment, in Fox, J D (Ed), *Language Testing Reconsidered*, Ottawa: University of Ottawa Press, 41–71.

Ballard, L and Lee, S (2015) *How young children respond to computerized reading and speaking test tasks*, paper presented at Language Testing Research Colloquium, Toronto, Canada, 18–20 March.

Bao, M, Egi, T and Han, Y (2011) Classroom study on noticing and recast features: capturing learner noticing with uptake and stimulated recall, *System* 39 (2), 215–228.

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

Bisson, M J, van Heuven, W J B, Conklin, K and Tunney, R J (2014) Processing of native and foreign language subtitles in films: an eye-tracking study, *Applied Psycholinguistics* 35 (2), 399–418.

Brunfaut, T and McCray, G (2015) *Looking into test-takers' cognitive processes whilst completing reading tasks: A mixed-method eye-tracking and stimulated recall study*, ARAGs Research Reports Online, Volume AR/2015/001, London: British Council, available online: www.britishcouncil.org/sites/default/files/brunfaut-and-mccray-report_final.pdf

Centre for Canadian Language Benchmarks (2012) *Canadian Language Benchmarks: English as a Second Language for Adults*, Ottawa: Centre for Canadian Language Benchmarks, available online: www.cic.gc.ca/english/pdf/pub/language-benchmarks.pdf

Chapelle, C A, Cotos, E and Lee, J (2015) Validity arguments for diagnostic assessment using automated writing evaluation, *Language Testing* 32 (3), 385–405.

Duchowski, A (2007) *Eye-tracking Methodology: Theory and Practice*, London: Springer.

Egi, T (2008) Investigating stimulated recall as a cognitive measure: Reactivity and verbal reports in SLA research methodology, *Language Awareness* 17 (3), 212–217.

Ertac, S (2011) Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback, *Journal of Economic Behavior & Organization* 80 (3), 532–545.

Fernández-Toro, M and Hurd, S (2014) A model of factors affecting independent learners' engagement with feedback on language learning tasks, *Distance Education* 35 (1), 106–125.

Fontanini, I and Braga Tomitch, L M (2009) Working memory capacity and L2 university students' comprehension of linear texts and hypertexts, *International Journal of English Studies* 9 (2), 1–18.

Gass, S M and Mackey, A (2016) *Stimulated Recall Methodology in Applied Linguistics and L2 Research*, New York: Routledge.

Godfroid, A and Schmidtke, J (2013) *What Do Eye Movements Tell Us About Awareness? A Triangulation of Eye-movement Data, Verbal Reports and Vocabulary Learning Scores*, Honolulu: University of Hawai'i at Manoa National Foreign Language Resource Center.

Godfroid, A and Spino, L (2015) Reconceptualizing reactivity of think-alouds and eye tracking: absence of evidence is not evidence of absence, *Language Learning* 65 (4), 896–928.

Holmqvist, K, Nyström, M, Andersson, R, Dewhurst, R, Jarodzka, H and van de Weijer, J (2011) *Eye-tracking: A Comprehensive Guide to Methods and Measures*, Oxford: Oxford University Press.

Jang, E E, Dunlop, M, Park, G and van der Boom, E H (2015) How do young students with different profiles of reading skill mastery, perceived ability, and goal orientation respond to holistic diagnostic feedback?, *Language Testing* 32 (3), 359–383.

Jones, N and Saville, N (2016) *Learning Oriented Assessment: A Systemic Approach*, Studies in Language Testing Volume 45, Cambridge: UCLES/ Cambridge University Press.

Just, M A and Carpenter, P A (1976) Eye fixations and cognitive processes, *Cognitive Psychology* 8 (4), 441–480.

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

Kang, H (2014) Understanding online reading through the eyes of first and second language readers: an exploratory study, *Computers & Education* 73, 1–8.

Kissler, J, Herbert, C, Winkler, I and Junghofer, M (2009) Emotion and attention in visual word processing: an ERP study, *Biological Psychology* 80 (1), 75–83.

Lam, W Y K (2008) Metacognitive strategy use: Accessing ESL learners' inner voices via stimulated recall, *Innovation in Language Learning and Teaching* 2 (3), 207–223.

Owen, N (2016) *An evidence-centred approach to reverse engineering: Comparative analysis of IELTS and TOEFL iBT reading sections*, unpublished doctoral thesis, University of Leicester.

Pintrich, P R (2004) A conceptual framework for assessing motivation and self-regulated learning in college students, *Educational Psychology Review* 16 (4), 385–407.

Posner, M I (1980) Orienting of attention, *Quarterly Journal of Experimental Psychology* 32 (1), 3–25.

Rayner, K (1998) Eye movements in reading and information processing: 20 years of research, *Psychological Bulletin* 124 (3), 372–422.

Richardson, D C, Dale, R and Spivey, M J (2007) Eye movements in language and cognition: A brief introduction, in Gonzalez-Marquez, M, Mittelberg, I, Coulson, S and Spivey, M J (Eds) *Methods in Cognitive Linguistics*, Amsterdam: John Benjamins Publishing Company, 323–344.

Segalowitz, N and Frenkiel-Fishman, S (2005) Attention control and ability level in a complex cognitive skill: Attention shifting and second language proficiency, *Memory & Cognition* 33 (4), 644–653.

Smith, B (2012) Eye tracking as a measure of noticing: A study of explicit recasts in SCMC, *Language Learning and Technology* 16 (3), 53–81.

Smither, J W, Brett, J F and Atwater, L E (2008) What do leaders recall about their multisource feedback?, *Journal of Leadership & Organizational Studies* 14 (3), 202–218.

Suvorov, R (2015) The use of eye tracking in research on video-based second language (L2) listening assessment: a comparison of context videos and content videos, *Language Testing* 32 (4), 463–483.

Sweller, J, Ayres, P and Kalyuga, S (2011) *Cognitive Load Theory*, New York: Springer.
Uusberg, A, Uibo, H, Kreegipuu, K and Tamm, M (2013) Unintentionality of affective attention across visual processing stages, *Frontiers in Psychology* 4, 969–970.
Wadlinger, H A and Isaacowitz, D M (2011) Fixing our focus: training attention to regulate emotion, *Personality and Social Psychology Review* 15 (1), 75–43.
Wagner, M (2015) *The centrality of cognitively diagnostic assessment for advancing secondary school ESL students' writing: A mixed methods study*, unpublished doctoral dissertation, University of Toronto.
Winke, P, Sydorenko, T and Gass, S (2013) Factors influencing the use of captions by foreign language learners: an eye-tracking study, *Modern Language Journal* 97 (1), 254–275.
Zhao, H (2010) Investigating learners' use and understanding of peer and teacher feedback on writing: a comparative study in a Chinese English writing classroom, *Assessing Writing* 15, 3–17.

8 Eye tracking and EEG in language assessment

Elaine Schmidt
Cambridge University Press & Assessment, UK

Carla Pastorino-Campos
Cambridge University Press & Assessment, UK

Abstract

Research in language assessment has largely concentrated on analysing output, i.e., learners' writing or speech, or results of multiple-choice or written answers in listening and reading tasks. While discussions of cognitive validity have become more frequent over the last few years, actual experimental data looking at cognitive processing during language assessment are sparse. Eye tracking and electroencephalography (EEG) can be used to tap into underlying cognitive processes which questionnaires and other written or oral responses cannot capture since they are by necessity influenced by deliberation. However, the few studies which have used eye-tracking tools in assessment contexts often analysed the data in a qualitative way by looking at retrospective think-aloud reports from participants. Thus, despite using a tool to investigate top-down cognitive processing, many studies have focused on using eye tracking to analyse subjective deliberations and attention-driven bottom-up processes. Additionally, no studies have used EEG in assessment yet. This chapter provides an overview of eye tracking and EEG methodologies, explaining which measures show which type of processing, and demonstrating the value they have added to research in language processing and learning in second language learners. Finally, we will briefly discuss the prospects of using each methodology in a language assessment context, specifically in the evaluation of processing load and in the development of test materials.

Introduction

Cognitive processing has become a buzzword in language assessment over the last few years. Despite its perceived popularity, relatively little research has been carried out that has looked at cognitive processing in language assessment contexts. In part, this can be explained by the use of terminology such as *cognitive validity* which is described as 'the extent to which the tasks

employed succeed in eliciting from candidates a set of processes which resemble those employed in a real-world speaking event' (Field 2011:65). However, while the question of whether a task reflects real-world processing should be of concern to us, it is not truly a matter of cognitive validity, but of ecological validity (see also Spinner, Gass and Behney (2013) for a study of ecological validity in eye tracking). Cognitive validity should concern itself with whether a task really tests the cognitive processing it claims to test. For example, does a listening task which employs pictures as answers really only test language comprehension, or does it also test visual processing? And is the level of processing difficulty adequate?

Cognitive processing has long been a focal point in linguistics and psychology, and has been tested extensively with online (i.e., real-time processing) tools such as eye tracking and EEG. In this chapter, we will provide a brief introduction to both tools. We will then demonstrate how they have contributed to research in linguistics and psychology and deepened our understanding of various aspects of language comprehension processes, the results of which are vital for language assessment research. We will also briefly discuss some of the studies that have used eye tracking in assessment but have not necessarily looked at cognitive processing. Finally, we will demonstrate how eye tracking and EEG can be usefully employed in assessment research in future.

Eye tracking: The background

Eye tracking is a non-invasive method in which participants' eye movements are recorded as they carry out a task. The tool shows processing in real time with a very high temporal resolution, with a time lag between the initiation of an eye movement and the execution of said movement of 150–175 milliseconds (ms) (Rayner, Slowiaczek, Clifton and Bertera 1983). While the eye covers a wide angle of the visual field (200 degrees), only a very small angle actually provides the brain with detailed information (two degrees). This area is called the fovea (Liversedge and Findlay 2000, Richardson, Dale and Spivey 2007). Eye movements are the result of two processes: bottom-up perception of the environment and top-down cognitive processes (Richardson et al 2007). Looking at objects because they are larger, or more saliently placed, reflects bottom-up perceptual processing, but not necessarily top-down cognitive processing. Top-down cognitive processing is often modulated by the difficulty of an item; for example, the time it takes to retrieve the meaning of said item, or to integrate the meaning into a specific context. Eye tracking can show cognitive processes because looking at an object shows what we are thinking of, and thus the time spent looking reflects the time thinking about this object (Carpenter and Just 1983, Tanenhaus, Spivey-Knowlton, Eberhard and Sedivy 1995). Consequently, the initial and total time spent

looking at an object (word, image, etc.) is a proxy for the difficulty the person experiences in processing the meaning of this object in the context they have been given (Kliegl, Nuthmann and Engbert 2006, Rayner 1998, Rayner and Reingold 2015, Reichle, Pollatsek, Fisher and Rayner 1998, Starr and Rayner 2001). Additionally, eye movements vary as a function of the syntactic and conceptual difficulty level of texts (Richardson et al 2007). Nevertheless, it should be noted that eye movements differ depending on the nature of the task, and vary depending on the input language. Eye movements in Chinese, for example, often differ due to the use of characters and the absence of spaces between words (Ma, Li, Xu and Li 2019).

While eye tracking is a relatively new addition in the field of assessment, the earliest eye-tracking research dates back to 1879 and has been successfully used to research cognitive processes for decades. The first era of eye tracking research (Rayner 1998) looked at the basic properties of eye movements, such as the fact that the eyes do not move smoothly across a text but jump (called *saccades*), that the average saccade length in reading is around seven to nine letters long, that reading is not a completely linear process but that readers frequently go back and re-read earlier parts of a sentence (*regressions*, which happen around 10–15% of the time while reading), that the average duration of a fixation is 200–250ms for skilled readers, or that the visual field is around 200 degrees, though the human eye only receives detailed information within around two degrees (Rayner 1998, Starr and Rayner 2001). The second era of eye-tracking research (between 1920 and 1970 according to Rayner 1998) focused on surface aspects of the tasks, i.e., on attentional aspects related to visual processing, such as which part of a screen participants look at when they complete a task. *Attention* is often described as either representing top-down (a.k.a. voluntary attention or endogenous attention) or bottom-up (a.k.a. reflexive attention or exogenous attention) processes. Top-down attention is directed by the intention and mental representations of behavioural goals, and interacts with the saliency-driven bottom-up attention. Second-era studies investigated bottom-up attentional processes (and, henceforth, this is how attention will be defined in this article). Considering this, while attention is one aspect of cognitive processing, these types of experiments do not show underlying cognitive processing *per se* since top-down processing is not necessarily required. It is only in the third era of eye-tracking research from around 1975 onwards that there was a move from just observing surface aspects to inferring underlying cognitive processes (Rayner 1998) in both reading and listening research.

Listening research with eye tracking capitalises on the fact that humans automatically look at an object that is mentioned (i.e., the word 'dog' will elicit looks to the image of a dog), even if participants have not been instructed to do so, as shown in studies using the visual world paradigm (see Huettig, Rommers and Meyer (2011) for a review). Importantly, the

object does not need to be specifically named, as a semantic activation of its concept is enough (i.e., 'a furry pet' will also lead to participants looking at the image of a dog). Eye movements even reveal listening comprehension when images are absent. When presented with auditory descriptions of events happening on different floors of skyscrapers, eye movements showed that participants fixated on a high point on a white screen when listening to an event that happened on the 29th floor, and moved their gaze downwards for events that were described on the 10th floor, even though the screen was completely white throughout the whole trial (Spivey and Geng 2001). However, participants are generally unaware of the eye movements or regressions they make.

> Participants typically denied looking to [the object] and yet their eye movements revealed a process substantially different from their conscious report and their manual action (Richardson et al 2007:337–338).

This shows that participants are often not just unaware of their cognitive processes; they nonetheless have strong convictions of what they believe they did. It is the tapping into underlying processes which led eye tracking to be widely adopted in the study of language, since behavioural responses rely on some meta-awareness of the participant and are skewed by subjective judgement. Another convincing example demonstrating the difference between behaviourally collected data and eye tracking tapping into subconscious processes comes from studies of *categorical perception*. Behavioural studies show that phonemes (for example /p/ and /b/) are distinguished in a categorical fashion, even if the sounds are presented in a gradient continuous manner. However, eye-tracking studies have demonstrated that even though the behavioural responses are categorical, and participants claim not to hear any fine-grained differences, their eye movements show uncertainties around category boundaries, demonstrating that the underlying subconscious processing is gradient (McMurray, Tanenhaus, Aslin and Spivey 2003). Eye tracking thus allows researchers to study top-down underlying cognitive processing precisely because eye movements are often automatic and often subconscious. As Richardson et al (2007:327) point out: 'briefly partially-active representations – that might never elicit reaching, speaking, or even internal monolog activity because they fade before reaching those thresholds – can nonetheless occasionally trigger an eye movement that betrays this otherwise-latent momentary consideration.' The information obtained from eye-tracking research is therefore free from subjective perceptions of participants that are abundant in questionnaires, where answers are by necessity influenced by deliberation. Consequently, behavioural data in the form of questionnaires, structured recall or similar can never reveal true underlying cognitive processes because

the processes are by their very nature subconscious. This is why studies in linguistics and psychology, where eye tracking has been used for decades, do not employ questionnaires or interviews about subjects' eye movements, and instead focus solely on the objective and quantifiable data which eye tracking provides. In addition to providing an unbiased view of cognitive processing, eye tracking can also give us insight into cognition as it unfolds in real time, thus complementing the picture of the outcomes of cognitive processing that traditional behavioural experiments provide.

Technical aspects

Eye tracking captures *fixations*. Fixations are moments when the eye remains 'still' looking at a specific object, i.e., not moving to another object/word (but note that the eyes never actually remain completely still; even in fixations there is a constant tremor which is called nystagmus (Rayner 1998)). The eye receives information from the centre of a fixation (fovea) and information that lies more than two degrees outside of the fovea (parafovea) (Richardson et al 2007). In order to measure cognitive processing, so-called Areas of Interest (AOIs) need to be defined on the text or image. These AOIs can be the length of a single word, a phrase (verb or noun phrase), a whole sentence, etc. In pictures, AOIs are typically defined in terms of pixels because the size of areas is less well defined than in text, where the number of letters can be a criterion for the size. AOIs allow researchers to establish whether a participant fixated on a particular area, and how often and how long they fixated on it. Based on the looking behaviour in those AOIs, researchers can then draw conclusions regarding processing difficulties. Typical measurements evaluating linguistic processing include first fixation duration, first pass time, saccades, number of regressions, and total time. The task type will usually determine which measures are appropriate. In the following, we will focus on the aforementioned measures only, describe what these terms mean, and what aspects of cognitive processing they measure.

First fixation duration. This is the duration of the first fixation on a word or on an object. First fixation duration measures initial parsing difficulties. As such, low-frequency words have longer first fixation durations than high-frequency words (Inhoff and Rayner 1986). Unexpected words, for example when they occur in a highly predictive context, also elicit longer first fixation durations as do longer words (with more letters) and content words in comparison to function words (Hyönä, Niemi and Underwood 1989, Inhoff 1984). In fact, short words, function words and highly predictable words are often skipped by skilled readers and not fixated on at all. The probability of a fixation is around 85% for a content word but only 35% for a function word (Carpenter and Just 1983). Generally, the complexity

of a word, text, or image determines first fixation durations (Jacobson and Dodwell 1979, Rayner 1998, Rayner and Pollatsek 1989). Thus, the more complex a word/object is, the longer the first fixation duration. Complexity can occur in different forms, from orthographic anomalies, to morphological irregularities, to frequency effects. First fixation duration is typically seen as a proxy for lexical access, as this constitutes the earliest reading process. A lower-frequency word is, for example, slower to be accessed in the mental lexicon, which is reflected in first fixation durations.

First pass/First pass time. First pass refers to the number of fixations a word/object elicits before the eye moves to the next item, whether it is to the left or right. First pass time, sometimes also called gaze duration, can be the same duration as a first fixation duration but does not necessarily have to be. Thus, if a word only elicits one fixation, then the first fixation duration and first pass time are the same. However, if a word elicits more than one fixation in the first go, the first pass time is the accumulation of the total time of fixation before the eyes move to the next word/object. First pass time is thus typically related to the length of a word; the longer a word, the more likely it is that a reader needs to fixate on multiple locations (Hyönä et al 1989). Additionally, violations or unexpected words often lead to longer first pass time since participants will likely fixate on words/objects multiple times (Rayner, Warren, Juhasz and Liversedge 2004).

Saccades. Saccades are forward movements of the eyes which are necessary so that the object that is focused on remains on the retina for effective processing (Starr and Rayner 2001). In a text that is read left to right, the eye will make saccades from each fixation to fixate on the next word. A saccade does not however need to be from one word to the next; it can also occur within the same word if the word is fixated multiple times on different locations. Saccades also occur on pictures/objects when the eye moves on to fixate on a different point in the picture. The motions of saccades are not smooth but little jumps, even though the eyes themselves remain stable (Starr and Rayner 2001). The average saccade length in reading has an angle of around two degrees, and a length of around seven to nine letters, but the actual size of the saccade depends on the spacing of the letters, the predictability of the text, the content of words, and the ability of readers (Richardson et al 2007). A string of letters with a high predictability, for example, typically elicits longer saccades while the length of saccades in complex texts decreases (Jacobson and Dodwell 1979, Rayner 1998, Rayner and Pollatsek 1989). Equally, words with low content such as function words are often not fixated on, and the saccade simply skips the function word. Poor readers typically have shorter saccades, so saccade length is a good indicator of reading ability (Starr and Rayner 2001).

Regressions. This term refers to backwards movements of the eye to where previous fixations have landed. Regressions typically occur in reading, even though readers are often unaware of them (Reichle et al 1998), and skilled readers perform regressions between 10% and 15% of their reading time. However, regressions are also intricately linked with reading ability and general cognitive abilities. Lower working memory, for example, leads to a larger number of regressions (as do poor reading skills) because the information is not sufficiently stored to access the meaning of the sentence without having to check that the correct interpretation was chosen (see Rayner (1985) for a link between eye movements and reading proficiency). Additionally, complexity affects the number of regressions in that more complex sentences lead to a larger number of regressions (Jacobson and Dodwell 1979, Rayner 1998, Rayner and Pollatsek 1989, Richardson et al 2007). Garden-path sentences, for example, typically lead to regressions because the initial interpretation of the sentence turns out to be incompatible with the information provided at the end of a sentence. Low-frequency words are also more likely to be re-fixated in a regression (Rayner and Duffy 1986, Rayner and Raney 1996). The number of regressions a reader makes can thus be used to determine processing difficulty and as a proxy for general reading difficulties, amongst others, since many regressions are simply the result of comprehension failures (Blanchard and Iran-Nejad 1987, Ehrlich and Rayner 1983, Frazier and Rayner 1982, Rayner 1998).

Total time. The total time in eye tracking refers to the amount of time a participant spent looking at an AOI during the whole trial. Consequently, it is an aggregation of first pass time and all other fixation times during subsequent visits (regressions) to the AOI. Total time thus reflects late processing and the integration-of-information stage where readers might re-evaluate or confirm the initial interpretation of a sentence (Starr and Rayner 2001). This means that total time is a good proxy for late stage processing difficulty since more difficult areas of the sentence are typically revisited more often, and more time is spent in those AOIs throughout the whole trial, in order to make sense of the sentence.

Electroencephalography (EEG)

Electroencephalography refers to the measuring of electrical activity from the cortex that are the result of 'communication' between synapses. The method was developed by German scientist Hans Berger and first applied to human subjects in 1929 to study their brain activity (La Vaque 1999). EEG is a non-invasive online tool which measures real-time processing at very high temporal resolutions. While it is possible to establish the source of neural activity to some level (i.e., seeing hemispheric differences or determining

whether activity is for example frontal vs. occipital, etc.), it has a relatively poor spatial resolution. Unlike positron emission tomography (PET), functional magnetic resonance imaging (fMRI), or magnetoencephalography (MEG) scans, it can thus not be used to determine the precise brain area that is associated with specific cognitive processes. In EEG, the electrical activity generated by the brain is captured by electrodes that are placed on the scalp, and saline gel is typically applied to establish connectivity.

In EEG research, the focus is often on event-related potentials (ERPs). ERPs are brain responses (also called *components*) to a specific trigger at a defined point in time. For example, if a participant encounters a semantic violation in a sentence (*For breakfast I always eat eggs and* ***pans***), then the ERP is measured from the moment the semantic violation (*pans*) is encountered. Crucially, a non-violation condition is required for comparison (*For breakfast I always eat eggs and* ***bacon***), and participants' responses to violations are computed in relation to non-violations in order to establish whether and how the brain has reacted to these semantic errors. The word trigger is typically used to refer to the time point at which an ERP is measured, i.e., the semantic violation is expected to trigger a response. In the example above, participants would typically show more negative activity in response to violations than non-violations, around 400ms after the onset of the violation. The timeline of ERPs (often indicated by the number in the name of the component, e.g., 400ms after stimulus onset in the N400), their directionality (positive vs. negative, indicated by a P or N in the name of the component), and their distribution on the scalp (e.g., left frontal) provide insights into the type of processing difficulty that is encountered. For example, while semantic violations trigger a negative activity 400ms after stimulus onset in the centro-parietal area, syntactic violations trigger a positive activity 600ms after onset. Note that there does not necessarily have to be a violation to trigger such a response. A semantically plausible word which is highly unlikely in a given context (*For breakfast I always eat eggs and* ***oysters***) would trigger the same type of response albeit with a smaller amplitude (i.e., the height of the peak of the curve). A lower amplitude thus results in a lower peak.

In the following we will describe the most common components in language-related research and which type of linguistic processing they are associated with.

Mismatch negativity (MMN). A mismatch negativity is a negative-going waveform which occurs roughly between 100 and 250ms after stimulus onset with a fronto-central distribution (Näätänen, Gaillard and Mäntysalo 1978, Pulvermüller et al 2001). As the name suggests, it is a response to a mismatch, in this instance of phones which in the case of adults tend to be phonemes. MMNs are typically elicited in oddball paradigms. In this

paradigm participants are exposed to a repetition of the same sound with the occasional irregular insertion of a different sound, e.g., /p/, /p/, /p/, /b/, /p/, /p/, /p/, etc. If a listener notices (even subconsciously) that a change in sound has occurred (/b/ instead of /p/), they will show an MMN. The MMN is thus frequently used as a test of categorical perception and could be used to test phoneme inventories in second language learners. Thus, if a learner does not have a distinction between the two sounds in their native language, they would not show an MMN. Equally, if they have not acquired the distinction in their L2, they would also not show an MMN.

N400. The N400 is a negative-going waveform between 370 and 520ms with a peak of 400ms after stimulus onset. It is typically right-lateralised and has a centro-parietal distribution (Chwilla, Brown and Hagoort 1995, Kutas and Hillyard 1980). It can be triggered through auditory or written input. When the N400 was first described (Kutas and Hillyard 1980), it was thought that it only occurs as a result of semantic violations. However, subsequent research has demonstrated that semantically unexpected or low-frequency words also trigger an N400, even though the amplitude of this N400 is lower than those in response to violations. If a participant encounters a low-frequency unexpected word in a highly predictive context, they need to revise their expectations to match the input. It is this process that shows as an N400. The N400 has thus been reinterpreted as a component indicating semantic re-integration (Friederici, Gunter, Hahne and Mauth 2004, Hagoort, Hald, Bastiaansen and Petersson 2004, Kutas and Federmeier 2011, Van Berkum 2009), which can be tested to determine familiarity with words in L2.

Left anterior negativity (LAN). The LAN is a negative-going waveform observed 300 to 500ms after stimulus onset in left-biased anterior regions (Friederici, Hahne and Mecklinger 1996). It is typically found in response to syntactic violations, such as disagreements in gender and number (Molinaro, Barber and Carreiras 2011). The LAN has been interpreted as reflecting the early detection of grammatical mismatches, particularly when followed by a P600 effect (see the next paragraph) in a biphasic pattern (Molinaro et al 2011; but see Tanner (2015) for a different view) or the increased working memory load presented by demanding language processing activities, such as the parsing of long-distance dependencies (Fiebach, Schlesewsky and Friederici 2002, Kluender and Kutas 1993). Together with the P600, the LAN can be considered as an indicator of syntactic complexity.

P600. The P600 is a positive-going waveform around 600ms after stimulus onset in midline channels over posterior regions. It is found in response to syntactic violations such as an inverse in word order (*The woman heard John's of account the story*) or garden-path sentences (Friederici et al 2004,

Osterhout and Holcomb 1992, Osterhout, Holcomb and Swinney 1994). Garden-path sentences contain temporary ambiguities which send the reader (less often the listener due to the disambiguating presence of prosody) down the wrong processing path. When new information towards the end of the sentence creates an incongruent interpretation with the initial parsing of the sentence, readers need to reanalyse the beginning of the sentence. For example, in the sentence '*the author wrote the novel was a success*' the noun phrase '*the novel*' is initially parsed as the object to the ambitransitive verb '*wrote*', and it is only when the reader encounters '*was*' instead of a new noun phrase that the initial parsing strategy fails and the sentence requires reanalysis. Thus, similar to the N400, the P600 can occur in response to violations, but can equally just occur as a result of complex or mis-parsed syntactic structures (Friederici et al 2004). The P600 is therefore a good indicator of how complex syntactic structures are in L2.

Insights from eye tracking and EEG in language learning and processing

As mentioned above, both eye tracking and EEG have been employed in the fields of linguistics and psychology for decades. In the following, we will highlight the most important findings in the L2 learning and processing field that have come from eye tracking and EEG research. We believe the results of these studies especially should be widely known by anyone doing research on cognitive processing in the language assessment field since they show the importance of language background and proficiency on processing, which need to be taken into consideration in the experimental design.

Eye tracking

Language processing. Eye tracking has been successfully employed to explore some of the crucial questions related to the acquisition and processing of grammar and words in second language learners. One important issue eye-tracking studies have helped elucidate is the extent to which non-native readers' grammatical knowledge and processing can be native-like in real time and, if not possible, where the differences lie. To do this, researchers have compared the eye movements of native and non-native readers when resolving language ambiguities or incongruences in sentence reading (for reviews, see Keating and Jegerski 2015, Roberts and Siyanova-Chanturia 2013, Siyanova-Chanturia 2013), as this comparison can provide us with information regarding how much native and non-native reading patterns overlap and, from this, infer whether learners' cognitive processing is similar to that of native speakers. Some of these studies have looked at syntactic ambiguity resolution, i.e., cases where a grammatically ambiguous sentence

could be resolved differently in a reader's L1 and L2 (Dussias and Sagarra 2007, Frenck-Mestre and Pynte 1997, Roberts, Gullberg and Indefrey 2008). For example, Frenck-Mestre and Pynte (1997) investigated native and non-native readers' grammatical ambiguity resolution strategies when presented with sentences in which the verb's attachment preferences could influence parsing. A sentence like '*Brutus hit the gladiator with the shield with his bared hands*' is ambiguous, as it can be interpreted as '*Brutus used the shield to hit*' (the prepositional phrase '*with the shield*' attached to the verb '*hit*') or as '*the gladiator had a shield*' (the prepositional phrase '*with the shield*' attached to the noun '*gladiator*').

The eye-tracking results showed that both native and non-native readers were able to make use of lexical information to disambiguate ambiguous sentences, as evidenced in late eye-tracking measures, but not in first fixation durations. Interestingly, in a second experiment, the researchers showed the effect of attachment mismatches between L1 and L2 (i.e., in cases where a verb had different attachment preferences in each language) which affected L2 readers in early processing even though they were subsequently able to resolve the ambiguity like native speakers, as evidenced by longer gaze durations in the critical area.

Eye tracking has also been able to demonstrate differences between native and non-native speakers in the detection and anticipation of morphological features (e.g., gender, see Dussias, Valdés Kroff, Guzzardo Tamargo and Gerfen 2013, Grüter, Lew-Williams and Fernald 2012) and the resolution of morphological disagreements (e.g., adverb-verbal inflection mismatch, see Ellis and Sagarra 2010, Sagarra and Ellis 2013). Ellis and Sagarra (2010) found that English native speakers and English-Spanish bilinguals of intermediate proficiency, whose L1 is not very rich in terms of morphological inflections, showed longer total time durations at the adverb (e.g., *today*) when presented with incongruent sentences (e.g., *today rained*). Instead, Spanish native speakers, whose language is morphologically richer, fixated longer on the verb (e.g., *today rained*). The authors consider this pattern as evidence that speakers and learners of a relatively poor morphological system prefer the adverb as a cue while speakers of a morphologically rich language prefer morphological cues. They also conclude that the attentional patterns learned for the L1, in this case the preference for lexical or morphological cues to infer time, have long-term effects, in some cases blocking the acquisition of cues that may be relevant, and indeed preferred, by other language systems.

Language selectivity questions, i.e., whether bi- or multilinguals store, activate and access their languages separately and selectively or whether they are integrated, have also been explored using eye tracking, most notably through the extensive body of work looking at ambiguous words such as cognates and interlingual homographs (e.g., Costa, Caramazza and Sebastian-Galles 2000, Degani and Tokowicz 2010, Duyck, Van Assche,

Drieghe and Hartsuiker 2007, Lemhöfer and Dijkstra 2004, Lemhöfer, Dijkstra and Michel 2004). For example, Blumenfeld and Marian (2007) recorded the eye movements of English-German and German-English bilinguals, and an English monolingual control group, as they heard words that were either cognates in their two languages, e.g., *arm* (English) vs. *Arm* (German), or control words, e.g., *bike* (English) vs. *Fahrrad* (German). The hypothesis was that, when listening to English words, the two groups of bilinguals (German L1 and English L1) would activate their German language representations, causing their gazes to divert to pictures of German competitors shown on a screen (i.e., the picture of an object whose German name is phonologically similar to the target) more than to controls (i.e., a picture of an object phonologically unrelated to the target). Indeed, this hypothesis was confirmed, as, when processing cognates, both groups of bilinguals looked at competitors more often than controls and were more likely to divert their gaze to the competitor than monolingual English speakers. When processing non-cognates, however, only German L1 speakers looked at the German competitors. These results led the researchers to conclude that high proficiency is required for the co-activation of non-target language representations when the words are not cognates, as only those bilinguals with higher German proficiency (German L1) gazed at German competitors when processing both cognates and non-cognates, whereas those with lower proficiency (English L1) did so only when the words were cognates.

Language learning. Eye tracking has also been used to inform some of the debates concerning language learning. Godfroid, Boers and Housen (2013), for example, found differences in reading times of words depending on their familiarity and whether they had been learned by the readers. In an incidental word learning design, advanced learners of English were exposed to sentences in which either a novel word (a pseudoword, e.g., *paniplines*), a known word (as a control, e.g., *boundaries*) or a combination of both (the known word serving as a cue for the meaning of the novel word, e.g., *boundaries-paniplines*, in both possible orders) were embedded. They also completed a vocabulary post-test to measure their retention of the novel words. As the researchers hypothesised, participants looked longer at unknown words than at known words in all measures. The addition of a semantic cue produced longer fixations in the known word if it followed the unknown word, but did not affect retention. On the other hand, longer total fixation times did result in an increased probability of recognising a novel word in the post-test. In other words, the longer a learner looked at an unknown word, the more chances of recognising it in the long term. Similar results were obtained when looking at the acquisition of irregular verb morphology (stem-vowel changes) in beginner learners of German, with longer fixations and a higher

number of comparisons between regular and irregular verbs, but not first fixation or gaze duration, related to positive learning effects (Godfroid and Uggen 2013). As exemplified by these two studies, eye tracking has been able to show that more time spent looking at novel linguistic information relates to better learning outcomes or, in other words, that attending to novel items is linked to higher long-term retention.

Language assessment. Unlike processing, and similar to learning, eye-tracking-based research on language testing is still in its infancy. The attempts at using eye tracking in this area have been primarily focused on establishing the cognitive validity of language tests. To achieve this goal, Bax (2013) used eye-tracking measures to observe the behaviour of Malaysian students with various language backgrounds as they were completing IELTS Reading test items. The items chosen for analysis were intended to capture careful and expeditious local reading processes. The results of the comparison of successful (i.e., those who responded correctly) to unsuccessful (i.e., those who responded incorrectly) students revealed significant differences in attention only for five of the 10 items analysed. The author interpreted these findings to suggest that successful students were more proficient at expeditious reading, spending less time on the text and locating and focusing on key information faster than unsuccessful candidates. While this study provides an interesting first look at reading behaviours in language testing, it had a number of limitations. First, the AOIs were not controlled for size. This means that longer fixations in larger areas are expected, regardless of the processing difficulty of the words or sentences themselves. Another crucial point to note is that the 38 students were 'randomly' sampled 'with first languages including Bahasa Melayu, Tamil, Chinese and others' (Bax 2013:447). This can be problematic because, as we have seen above, the L1 influences reading times (e.g., cognate effects, syntactic similarity). In addition, the learners had completely different scripts in their native language (as far as we know since not all L1s were actually listed). Tamil, for example, is an alphabetic language with consonants and vowels represented; however, the script is different to English. In contrast, Chinese is a logographic language. Do readers of these languages all show the same reading patterns? If languages were included where reading occurs from right to left, this would create additional confounds because attention would be distributed differently across the screen. The study then also compared students who answered correctly (successful) and incorrectly (unsuccessful) based on the attention paid to the text prompt, key sections of the text at a sentence level or larger and key sections of the text and question at a word or phrase level. Thus, instead of investigating the underlying top-down cognitive processes in reading with AOIs of comparable size, the study looked at attention-driven bottom-up processing. With these limitations in mind, and considering that

effects were found for only half of the items, it is difficult to draw any strong conclusions from this study. Furthermore, conflicting results were obtained using a similar design but testing students from another variety of linguistic backgrounds (Bax 2015). Despite having used the same materials, some of the items in which significant differences between successful and unsuccessful candidates were found before were not present in the second experiment. These divergent results were attributed to the different nationalities and linguistic backgrounds of the students in the two studies, a feature that was already present in the first study (albeit restricted to linguistic diversity).

A more recent study investigated the effect of test item types on candidates' reading processes (Bax and Chan 2019). As in previous studies, successful and unsuccessful candidates' eye-tracking measures (fixation duration and number and visit duration and number) were compared and an additional qualitative visual analysis of gaze patterns was conducted. Based on these analyses, the authors concluded that the type of item (cloze multiple choice, multiple-choice question, cloze, summary cloze, heading matching, which texts) influenced the type of reading process the student used, eliciting both lower- and higher-level reading processes and activities. As for other research reviewed in this section, this study had a number of drawbacks, including a small sample size, little information about the linguistic background of participants, and technical difficulties with the setting up of the eye-tracking experiment (e.g., differences in AOI size).

Compared with the studies reviewed on the language processing and learning sections, the studies focused on assessment were not as controlled in terms of the participants (number per group, language background), utilised less typical and less nuanced eye-tracking measures (e.g., no differentiation between first and subsequent gaze durations, different sizes in the areas of interest), and used a less sophisticated eye-tracking apparatus. Despite these limitations, these studies have paved the way for researchers to continue exploring and refining the ways in which eye-tracking technology can inform language testing. They highlight the need to look further afield and incorporate findings from areas that have carried out eye-tracking research for decades in order to arrive at meaningful conclusions. Further research investigating higher-order cognitive processing, and thus the cognitive validity of tests, as opposed to attention-driven processing, is needed, as the former is informative in the design of test content, while the latter can inform other aspects of test construction, such as layout.

EEG

As with eye tracking, EEG has been used to investigate native and non-native speakers' language processing and learning in order to determine whether learners can exhibit native-like neuro-cognitive processing.

Language processing. Studies using EEG have been instrumental in helping us understand how humans process sounds (MMN, Garrido, Kilner, Stephan and Friston 2009), words (N400, Kutas and Federmeier 2011) and morphosyntax (N400 and P600, Caffarra, Molinaro, Davidson and Carreiras 2015, Morgan-Short 2014) in a second language. For example, Gillon Dowens, Vergara, Barber and Carreiras (2010) investigated whether the ERPs signature of late but highly proficient and immersed Spanish learners (English L1) would be similar to those of native Spanish speakers when processing morphosyntactic violations. This is especially important because it shows whether highly proficient learners can indeed process morphosyntax like native speakers. Of particular interest was to observe whether the overlap of syntactic features between the learners' L1 and L2 would have an effect on their processing in real time. To do this, participants' electrophysiological responses were recorded as they read sentences that were either correct or contained one of two types of morphosyntactic agreement violations, gender or number. Both English (L1) and Spanish (L2) instantiate number agreement between articles and nouns, but gender-based agreement is only present in Spanish (L2). For this reason, if L1-L2 similarity influences ERPs when reading in the L2, it should be expected that there would be differences between responses to violations of a feature present in both L1 and L2 (number) and a feature unique to the L2 (gender). Native speakers of Spanish in this experiment showed the expected LAN-P600 diphasic pattern to violations of both number and gender, as well as an extended negativity between 1,000 and 1,300ms that was more negative for violations. The late L2 learners also showed a P600 and a late negativity for both gender and number violations and an additional early negativity only in those cases where the violation appeared in first-sentence position. Interestingly, contrary to native speakers, learners showed differences in the magnitude of the P600 and late negativity based on whether the item was a number or a gender violation, with stronger effects observed for number violations in both time windows. Furthermore, the early negativity observed for non-native speakers had a longer duration when processing number violations and had a later onset when processing gender violations when compared to native speakers. The authors conclude that, while similar electrophysiological patterns between native and highly proficient non-native speakers may be observed during morphosyntactic processing, some factors pertaining to the learners (e.g., age of acquisition) and to the languages involved (e.g., L1-L2 similarity) may moderate these similarities.

Some efforts have been made to explore these factors. As observed in eye-tracking research, proficiency has been shown to be a very important variable to control for or, indeed, to focus on when attempting to determine whether learners can exhibit native-like neurocognitive processing patterns. Bowden, Steinhauer, Sanz and Ullman (2013) found that, when being

exposed to sentences with syntactic (word order) violations, advanced Spanish learners exhibited similar LAN and P600 effects to those of native speakers, in contrast to a beginner L2 group. It is important to note that participants in this experiment were carefully selected to have different levels of proficiency, amount of classroom experience and study-abroad experience, but were matched with regard to age of acquisition or additional immersion experience (besides study-abroad). Similar immersion effects, but a surprisingly different effect of proficiency, were found by Tanner, Inoue and Osterhout (2014). In this study, results showed the N400-P600 effect in response to subject-verb agreement violations in a group of very proficient Spanish-English bilinguals. They noted, however, that there were not only large differences between participants in terms of the magnitude of the response, but also with regard to the components, with some showing primarily an N400 effect while others only displaying P600. Participants with more immersion experience (as gathered from an earlier age of arrival) and more motivation to speak in their L2 showed more P600 dominance, while those who were more proficient in their L2 (as measured by performance in a proficiency test) showed increased magnitude of response across both time windows, irrespective of the type of response.

With regard to word processing, decades of research involving native speakers has shown that N400 effects may be observed when participants process meaning (for a comprehensive review, see Kutas and Federmeier 2011). As with other aspects of language acquisition and processing, the question of whether the N400 effect would be similar between native and non-native speakers quickly arose, generating many insights but also new questions about the multilingual brain as researchers addressed each issue. In an early study, Hahne and Friederici (2001) recorded the ERPs of a group of 12 late learners of German (Japanese L1) with immersion experience as they heard German sentences that were either correct or that contained semantic, syntactic or both types of violations and compared them to L1s (Hahne and Friederici 2002). As expected, the learners exhibited a significant N400 effect when listening to sentences with semantic violations, which was indistinguishable in amplitude from that of the native speakers (Hahne and Friederici 2002), even though the time window was longer.

Contrary to the findings on syntactic processing, some initial evidence indicated that proficiency may not necessarily play a role in the N400 response of language learners as they process semantic stimuli. For example, Kotz and Elston-Güttler (2004) studied groups of German L1, and English L2 learners (high and low proficiency, based on differences in months of exposure to their L2 while abroad, self-assessed L2 proficiency, performance on a vocabulary test and on an L2 memory recall task). Their neural patterns were recorded while they read pairs of words that were either associated, i.e., the pair usually co-occur in language (e.g., flower-vase) or categorically

related, i.e., the pair share common characteristics or membership to a larger category (e.g., flower-rose). Both proficiency groups exhibited the expected N400 effect when processing associative pairs, but not when processing categorical pairs. This pattern of results is in contrast to those obtained with native speakers and early bilinguals (Kotz 2001), who showed N400 effects for both types of semantic relationships. The researchers interpret these findings as evidence for an age-of-acquisition influence on the strength of the word-to-word (associative) and word-to-concept (categorical) relationships formed in a bilingual lexicon, as no proficiency-based effects were found in this sample and, when compared to previous studies, these learners differed primarily on the age at which they had begun learning their L2, i.e., early in Kotz (2001) and late in Kotz and Elston-Güttler (2004).

EEG has also been used to explore non-native neural patterns when processing sounds in an L2. Dehaene-Lambertz (1997) showed that native French speakers only demonstrated a significant MMN effect to sound changes that represented consonant phonemes in their own phonological system, as opposed to changes that were not. Similarly, Näätänen et al (1997) found that native speakers of Finnish and Estonian showed increased MMN responses when exposed to vocalic contrasts which were also present in their L1. In contrast, Finnish speakers did not evidence MMN effects when processing an Estonian-only deviant. Early evidence suggested that long-term use and contact with an L2, possibly including immersion experience, may be necessary to develop native-like responses to sounds not present in a learner's L1 phonemic inventory. For example, Winkler et al (1999) found that only Hungarian-Finnish bilinguals with extensive immersion experience in their L2 (Finnish) showed similar MMN responses as Finnish native speakers to contrasts only present in that language. Additionally, Peltola et al (2003) showed that learning Finnish as an L2 in the classroom may not be sufficient to develop native-like MMN responses to vocalic contrasts. In their study, Finnish-English late bilinguals who were very proficient in their L2 had a significantly smaller MMN response in comparison to English-only controls despite their high level of proficiency. The magnitude of the ERP response was also modulated by acoustic distance for this group, with a more distant contrast (a more salient difference) eliciting larger responses.

Language learning. The results of longitudinal studies looking at the child and adult acquisition of L2 sounds in non-immersed environments challenge the idea that long-term exposure and immersion is necessary for learners to exhibit native-like neural patterns. For example, Cheour, Shestakova, Alku, Ceponiene and Näätänen (2002) found that Finnish L1 children who attended a school or nursery where French was spoken over 50% of the time showed a marked increase in MMN amplitude from the first measurement, when they joined the school or nursery, to the second measurement, two

months later. This indicates that children developed a sensibility for non-native (French) sounds within a matter of months, despite them not being completely immersed in the L2 and in the absence of direct instruction regarding the sounds. More strikingly, Tamminen, Peltola, Kujala and Näätänen (2015) found that when a group of Finnish adult native speakers were specifically taught to perceive a voicing contrast present in English but not in their L1, their electrophysiological response exhibited MMN effects after only two days of very brief training (of only a few minutes per day).

Changes in neural responses to grammatical violations were also observed after a relatively short period of intensive language instruction in White, Genesee and Steinhauer's (2012) longitudinal study. Their results showed that a P600 effect, not present at the beginning of the course, emerged when Korean and Chinese L1 participants were tested at the end, regardless of their L1. Furthermore, proficiency in the grammaticality judgement task, during which ERPs were measured, modulated the magnitude of the response, with higher scores associated with bigger P600 effects. These findings were interesting, as the L2 feature tested, English regular past tense, is not present in one group's L1 (Chinese, no inflectional morphology) and is present, but instantiated differently, in the other (Korean). Based on these results, White and colleagues conclude that learners' L1 may have a smaller role in the type of neural responses they may develop, as late learners in this experiment did show a P600 effect after intensive instruction that led to increased proficiency, regardless of whether transfer from the L1 would be possible (Chinese) or beneficial (Korean).

Eye tracking and EEG in language learning, processing and testing: Summary

The research described here provides a brief overview of the uses of eye tracking and EEG in the study of language learning, processing and testing. These studies have used many different designs, using sentences (Ellis and Sagarra 2010) and single-word stimuli (Kotz and Elston-Güttler 2004), presented either visually (Gillon Dowens et al 2010) or auditorily (Winkler et al 1999) and using established formats, such as the syntactic violation (Frenck-Mestre and Pynte 1997) or visual-world paradigm (Blumenfeld and Marian 2007) in eye tracking or the oddball paradigm (Dehaene-Lambertz 1997) in EEG. The moderating effects of factors pertaining to learners (e.g., proficiency, immersion experience, age of acquisition), their languages (e.g., similarity) have been manipulated in several instances to gain a deeper understanding of the phenomena studied. The effect of proficiency (Peltola et al 2003), language similarity (Gillon Dowens et al 2010) and immersion experience (Bowden et al 2013) have been particularly researched, with conflicting but very informative results.

Eye tracking and EEG data has helped inform debates on whether bilinguals or multilinguals selectively activate words in one or the other language, as required, and what might be the factors that affect this language selectivity (Blumenfeld and Marian 2007). By comparing native and non-native speakers, researchers have been able to observe whether learners' processing can be native-like when resolving syntactic ambiguity (Tanner et al 2014), reacting to grammatical gender mismatches (Gillon Dowens et al 2010), identifying lack of agreement between adverbs and morphological inflections (Ellis and Sagarra 2010), or when processing semantic violations (Hahne and Friederici 2002) or sounds in their L1 and L2 (Näätänen et al 1997). We have also gained more insights into how we learn new words (Godfroid et al 2013), morphological rules (Godfroid and Uggen 2013) and sounds (Cheour et al 2002), and how proficiency affects the learning of these. It is especially the ability to look at proficiency on a cognitive or neural processing level, rather than just behavioural output responses, that could be beneficial for assessment of the future. At a more applied level, eye tracking, but not EEG, is beginning to be used to obtain more evidence about the validity of language tests (Bax 2013) and the behaviour of candidates as they are being assessed (Bax 2015, Bax and Chan 2019). Part of the reason why EEG is used less frequently in assessment research is the time-consuming and technical nature of data analysis, as well as practical concerns such as placing electrodes with gel on participants' scalps which requires washing the participants' hair afterwards. This makes EEG less practical in settings without adequate facilities. However, the possibility of observing how learners process linguistic information in real time, as inferred by their eye movements and neural patterns, has allowed us to construct a clearer picture of language learning, processing and testing, and continues to be applied in more and more innovative ways. In the following section we will briefly discuss what the future holds for eye tracking and EEG in language assessment.

The future of eye tracking and EEG in language assessment

While a number of studies have already used eye tracking in assessment, these have almost exclusively focused on attention-driven processing, comparable to Era 2 of eye-tracking research. The future of eye-tracking research in assessment lies in entering Era 3, i.e., with the focus on the actual underlying cognitive processes. For example, eye tracking and EEG could be used to determine the level of difficulty of items. As shown by linguistics research, some questions need to be addressed on whether the level of difficulty in items is *appropriate for the level of proficiency*. While this is to a large extent already done for vocabulary, linguistic research shows that morphology,

syntax, and phonetics/phonology are other factors that need to be taken into consideration when designing reading or listening items. Difficulty can easily be measured with eye tracking while the cause for difficulty (i.e., whether it is rooted in syntax, morphology, etc.) can be tested with EEG. An even more important question is whether the *level of difficulty is comparable across language backgrounds* or whether we accidentally and perhaps even systematically disadvantage learners from certain language backgrounds due to the lexicon or syntactic structures we use when constructing the items. Both eye tracking and EEG can be used to answer this question. Recently, the co-registration of EEG and eye tracking (simultaneous recording of brain activity while capturing eye movements) has also gained traction in psychology and linguistics. This is another exciting area since the combination of the two methodologies can bring further as yet unexplored insights.

Going forward, it is also important to note that testing bottom-up attention-driven processing will necessarily require different criteria to top-down cognitive processing in the way eye-tracking experiments are designed. Items for the latter may not be presentable in the test format in which they would appear in the actual assessment, but may need to be changed in order to have comparable sizes of AOIs in order to draw meaningful conclusions; in order to determine differences in *cognition*, effects of *attention* need to be mitigated. EEG studies will need careful design as well in order to timelock to the correct part of the sentence. Crucially, the overview of the linguistic studies above has hopefully demonstrated the importance of controlling for language background when testing the actual cognitive processes in reading or listening comprehension. Numerous studies have shown the influence of L1 in language processing so ignoring language background or haphazardly mixing language backgrounds means that these effects are potentially washed out or artificially magnified. The same holds true for different proficiency levels. The only meaningful results in eye-tracking and EEG studies on cognitive processing come from studies which carefully control for these factors.

References

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

Bax, S (2015) *Using eye-tracking to research the cognitive processes of multinational readers during an IELTS reading test*, IELTS Research Reports Online Series.

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

Blanchard, H E and Iran-Nejad, A (1987) Comprehension processes and eye movement patterns in the reading of surprise-ending stories, *Discourse Processes* 10 (1), 127–138.

Blumenfeld, H K and Marian, V (2007) Constraints on parallel activation in bilingual spoken language processing: Examining proficiency and lexical status using eye-tracking, *Language and Cognitive Processes* 22 (5), 633–660.

Bowden, H W, Steinhauer, K, Sanz, C and Ullman, M T (2013) Native-like brain processing of syntax can be attained by university foreign language learners, *Neuropsychologia* 51 (13), 2,492–2,511.

Caffarra, S, Molinaro, N, Davidson, D and Carreiras, M (2015) Second language syntactic processing revealed through event-related potentials: An empirical review, *Neuroscience & Biobehavioral Reviews* 51, 31–47.

Carpenter, P A and Just, M A (1983) What your eyes do while your mind is reading, in Rayner, K (Ed) *Eye Movements in Reading: Perceptual and Language Processes*, New York: Academic Press, 275–307.

Cheour, M, Shestakova, A, Alku, P, Ceponiene, R and Näätänen, R (2002) Mismatch negativity shows that 3–6-year-old children can learn to discriminate non-native speech sounds within two months, *Neuroscience Letters* 325 (3), 187–190.

Chwilla, D J, Brown, C M and Hagoort, P (1995) The N400 as a function of the level of processing, *Psychophysiology* 32 (3), 274–285.

Costa, A, Caramazza, A and Sebastian-Galles, N (2000) The cognate facilitation effect: Implications for models of lexical access, *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26 (5), 1,283–1,296.

Degani, T and Tokowicz, N (2010) Semantic ambiguity within and across languages: An integrative review, *Quarterly Journal of Experimental Psychology* 63 (7), 1,266–1,303.

Dehaene-Lambertz, G (1997) Electrophysiological correlates of categorical phoneme perception in adults, *NeuroReport* 8 (4), 919–924.

Dussias, P E and Sagarra, N (2007) The effect of exposure on syntactic parsing in Spanish–English bilinguals, *Bilingualism: Language and Cognition* 10 (1), 101–116.

Dussias, P E, Valdés Kroff, J R, Guzzardo Tamargo, R E and Gerfen, C (2013) When Gender and Looking Go Hand in Hand: Grammatical Gender Processing In L2 Spanish, *Studies in Second Language Acquisition* 35 (2), 353–387.

Duyck, W, Van Assche, E, Drieghe, D and Hartsuiker, R J (2007) Visual word recognition by bilinguals in a sentence context: Evidence for nonselective lexical access, *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33 (4), 663–679.

Ehrlich, K and Rayner, K (1983) Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing, *Journal of Verbal Learning and Verbal Behavior* 22 (1), 75–87.

Ellis, N C and Sagarra, N (2010) Learned attention effects in L2 temporal reference: The first hour and the next eight semesters, *Language Learning* 60, 85–108.

Fiebach, C J, Schlesewsky, M and Friederici, A D (2002) Separating syntactic memory costs and syntactic integration costs during parsing: the processing of German WH-questions, *Journal of Memory and Language* 47 (2), 250–272.

Field, J (2011) Cognitive validity, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing Volume 30, Cambridge: UCLES/Cambridge University Press, 65–112.

Frazier, L and Rayner, K (1982) Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences, *Cognitive Psychology* 14 (2), 178–210.
Frenck-Mestre, C and Pynte, J (1997) Syntactic Ambiguity Resolution While Reading in Second and Native Languages, *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology* 50 (1), 119–148.
Friederici, A D, Hahne, A and Mecklinger, A (1996) Temporal structure of syntactic parsing: Early and late event-related brain potential effects, *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22 (5), 1,219–1,248.
Friederici, A D, Gunter, T C, Hahne, A and Mauth, K (2004) The relative timing of syntactic and semantic processes in sentence comprehension, *NeuroReport* 15 (1), 165–169.
Garrido, M I, Kilner, J M, Stephan, K E and Friston, K J (2009) The mismatch negativity: A review of underlying mechanisms, *Clinical Neurophysiology* 120 (3), 453–463.
Gillon Dowens, M, Vergara, M, Barber, H A and Carreiras, M (2010) Morphosyntactic Processing in Late Second-Language Learners, *Journal of Cognitive Neuroscience* 22 (8), 1,870–1,887.
Godfroid, A and Uggen, M S (2013) Attention to Irregular Verbs by Beginning Learners of German: An Eye-Movement Study, *Studies in Second Language Acquisition* 35 (2), 291–322.
Godfroid, A, Boers, F and Housen, A (2013) An Eye for Words: Gauging the Role of Attention in Incidental L2 Vocabulary Acquisition by Means of Eye-Tracking, *Studies in Second Language Acquisition* 35 (3), 483–517.
Grüter, T, Lew-Williams, C and Fernald, A (2012) Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research* 28 (2), 191–215.
Hagoort, P, Hald, L, Bastiaansen, M and Petersson, K M (2004) Integration of word meaning and world knowledge in language comprehension, *Science* 304 (5669), 438–441.
Hahne, A and Friederici, A D (2001) Processing a second language: late learners' comprehension mechanisms as revealed by event-related brain potentials, *Bilingualism: Language and Cognition* 4 (2), 123–141.
Hahne, A and Friederici, A D (2002) Differential task effects on semantic and syntactic processes as revealed by ERPs, *Cognitive Brain Research* 13 (3), 339–356.
Huettig, F, Rommers, J and Meyer, A S (2011) Using the visual world paradigm to study language processing: A review and critical evaluation, *Acta Psychologica* 137 (2), 151–171.
Hyönä, J, Niemi, P and Underwood, G (1989) Reading long words embedded in sentences: Informativeness of word halves affects eye movements, *Journal of Experimental Psychology: Human Perception and Performance* 15 (1), 142–152.
Inhoff, A W (1984) Two stages of word processing during eye fixations in the reading of prose, *Journal of Verbal Learning and Verbal Behavior* 23 (5), 612–624.
Inhoff, A W and Rayner, K (1986) Parafoveal word processing during eye fixations in reading: Effects of word frequency, *Perception & Psychophysics* 40 (6), 431–439.
Jacobson, Z and Dodwell, P C (1979) Saccadic eye movements during reading, *Brain and Language* 8, 303–314.

Keating, G D and Jegerski, J (2015) Experimental Designs in Sentence Processing Research: A Methodological Review and User's Guide, *Studies in Second Language Acquisition* 37 (1), 1–32.
Kliegl, R, Nuthmann, A and Engbert, R (2006) Tracking the mind during reading: The influence of past, present, and future words on fixation durations, *Journal of Experimental Psychology: General* 135 (1), 12–35.
Kluender, R and Kutas, M (1993) Bridging the Gap: Evidence from ERPs on the Processing of Unbounded Dependencies, *Journal of Cognitive Neuroscience* 5 (2), 196–214.
Kotz, S A (2001) Neurolinguistic evidence for bilingual language representation: a comparison of reaction times and event-related brain potentials, *Bilingualism: Language and Cognition* 4 (2), 143–154.
Kotz, S A and Elston-Güttler, K (2004) The role of proficiency on processing categorical and associative information in the L2 as revealed by reaction times and event-related brain potentials, *Journal of Neurolinguistics* 17 (2–3), 215–235.
Kutas, M and Federmeier, K D (2011) Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP), *Annual Review of Psychology* 62 (1), 621–647.
Kutas, M and Hillyard, S A (1980) Reading senseless sentences: Brain potentials reflect semantic incongruity, *Science* 207 (4427), 203–205.
La Vaque, T J (1999) The history of EEG Hans Berger: Psychophysiologist. A historical vignette, *Journal of Neurotherapy* 3 (2), 1–9.
Lemhöfer, K and Dijkstra, T (2004) Recognizing cognates and interlingual homographs: Effects of code similarity in language-specific and generalized lexical decision, *Memory & Cognition* 32 (4), 533–550.
Lemhöfer, K, Dijkstra, T and Michel, M (2004) Three languages, one ECHO: Cognate effects in trilingual word recognition, *Language and Cognitive Processes* 19 (5), 585–611.
Liversedge, S P and Findlay, J M (2000) Saccadic eye movements and cognition, *Trends in Cognitive Sciences* 4 (1), 6–14.
Ma, G, Li, Z, Xu, F and Li, X (2019) The modulation of eye movement control by word length in reading Chinese, *Quarterly Journal of Experimental Psychology* 72 (7), 1,620–1,631.
McMurray, B, Tanenhaus, M K, Aslin, R N and Spivey, M J (2003) Probabilistic constraint satisfaction at the lexical/phonetic interface: Evidence for gradient effects of within-category VOT on lexical access, *Journal of Psycholinguistic Research* 32 (1), 77–97.
Molinaro, N, Barber, H A and Carreiras, M (2011) Grammatical agreement processing in reading: ERP findings and future directions, *Cortex* 47 (8), 908–930.
Morgan-Short, K (2014) Electrophysiological Approaches to Understanding Second Language Acquisition: A Field Reaching its Potential, *Annual Review of Applied Linguistics* 34, 15–36.
Näätänen, R, Gaillard, A W and Mäntysalo, S (1978) Early selective-attention effect on evoked potential reinterpreted, *Acta Psychologica* 42 (4), 313–329.
Näätänen, R, Lehtokoski, A, Lennes, M, Cheour, M, Huotilainen, M, Iivonen, A, Vainio, M, Alku, P, Ilmoniemi, R J, Luuk, A, Allik, J, Sinkkonen, J and Alho, K (1997) Language-specific phoneme representations revealed by electric and magnetic brain responses, *Nature* 385, 432–434.
Osterhout, L and Holcomb, P J (1992) Event-related brain potentials elicited by syntactic anomaly, *Journal of Memory and Language* 31 (6), 785–806.

Osterhout, L, Holcomb, P J and Swinney, D A (1994) Brain potentials elicited by garden-path sentences: evidence of the application of verb information during parsing, *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20 (4), 786–803.

Peltola, M S, Kujala, T, Tuomainen, J, Ek, M, Aaltonen, O and Näätänen, R (2003) Native and foreign vowel discrimination as indexed by the mismatch negativity (MMN) response, *Neuroscience Letters* 352 (1), 25–28.

Pulvermüller, F, Kujala, T, Shtyrov, Y, Simola, J, Tiitinen, H, Alku, P, Alho, K, Martinkauppi, S, Ilmoniemi, R N and Näätänen, R (2001) Memory traces for words as revealed by the mismatch negativity, *Neuroimage* 14 (3), 607–616.

Rayner, K (1985) The role of eye movements in learning to read and reading disability, *Remedial and Special Education* 6 (6), 53–60.

Rayner, K (1998) Eye movements in reading and information processing: 20 years of research, *Psychological Bulletin* 124 (3), 372–422.

Rayner, K and Duffy, S A (1986) Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity, *Memory & Cognition* 14 (3), 191–201.

Rayner, K and Pollatsek, A (1989) *The Psychology of Reading*, New York: Erlbaum.

Rayner, K and Raney, G E (1996) Eye movement control in reading and visual search: Effects of word frequency, *Psychonomic Bulletin & Review* 3, 245–248.

Rayner, K and Reingold, E M (2015) Evidence for direct cognitive control of fixation durations during reading, *Current Opinion in Behavioral Sciences* 1, 107–112.

Rayner, K, Slowiaczek, M L, Clifton, C and Bertera, J H (1983) Latency of sequential eye movements: Implications for reading, *Journal of Experimental Psychology: Human Perception and Performance* 9 (6), 912–922.

Rayner, K, Warren, T, Juhasz, B J and Liversedge, S P (2004) The Effect of Plausibility on Eye Movements in Reading, *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30 (6), 1,290–1,301.

Reichle, E D, Pollatsek, A, Fisher, D L and Rayner, K (1998) Toward a model of eye movement control in reading, *Psychological Review* 105 (1), 125–157.

Richardson, D C, Dale, R and Spivey, M J (2007) Eye movements in language and cognition, in Gonzalez-Marquez, M, Mittelberg, I, Coulson, S and Spivey, M J (Eds) *Methods in Cognitive Linguistics*, Amsterdam: John Benjamins, 323–344.

Roberts, L and Siyanova-Chanturia, A (2013) Using Eye-tracking to Investigate Topics in L2 Acquisition and L2 Processing, *Studies in Second Language Acquisition* 35 (2), 213–235.

Roberts, L, Gullberg, M and Indefrey, P (2008) Online Pronoun Resolution in L2 Discourse: L1 Influence and General Learner Effects, *Studies in Second Language Acquisition* 30 (3), 333–357.

Sagarra, N and Ellis, N C (2013) From Seeing Adverbs to Seeing Verbal Morphology: Language Experience and Adult Acquisition of L2 Tense, *Studies in Second Language Acquisition* 35 (2), 261–290.

Siyanova-Chanturia, A (2013) Eye-tracking and ERPs in multi-word expression research: A state-of-the-art review of the method and findings, *The Mental Lexicon* 8 (2), 245–268.

Spinner, P, Gass, S and Behney, J (2013) Ecological Validity in Eye-Tracking: An Empirical Study, *Studies in Second Language Acquisition* 35 (2), 389–415.

Spivey, M J and Geng, J J (2001) Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects, *Psychological Research* 65 (4), 235–241.

Starr, M S and Rayner, K (2001) Eye movements during reading: Some current controversies, *Trends in Cognitive Sciences* 5 (4), 156–163.

Tamminen, H, Peltola, M S, Kujala, T and Näätänen, R (2015) Phonetic training and non-native speech perception — New memory traces evolve in just three days as indexed by the mismatch negativity (MMN) and behavioural measures, *International Journal of Psychophysiology* 97 (1), 23–29.

Tanenhaus, M K, Spivey-Knowlton, M J, Eberhard, K M and Sedivy, J C (1995) Integration of visual and linguistic information in spoken language comprehension, *Science* 268 (5217), 1,632–1,634.

Tanner, D (2015) On the left anterior negativity (LAN) in electrophysiological studies of morphosyntactic agreement: A Commentary on 'Grammatical agreement processing in reading: ERP findings and future directions' by Molinaro et al., 2014, *Cortex* 66, 149–155.

Tanner, D, Inoue, K and Osterhout, L (2014) Brain-based individual differences in online L2 grammatical comprehension, *Bilingualism: Language and Cognition* 17 (2), 277–293.

Van Berkum, J J (2009) The neuropragmatics of 'simple' utterance comprehension: An ERP review, in Sauerland, U and Yatsushiro, K (Eds) *Semantics and Pragmatics: From Experiment to Theory*, Basingstoke: Palgrave Macmillan, 276–316.

White, E J, Genesee, F and Steinhauer, K (2012) Brain Responses Before and After Intensive Second Language Learning: Proficiency Based Changes and First Language Background Effects in Adult Learners, *PLOS ONE* 7 (12), e52318.

Winkler, I, Kujala, T, Tiitinen, H, Sivonen, P, Alku, P, Lehtokoski, A, Czigler, I, Csepe, V, Ilmoniemi, R J and Näätänen, R (1999) Brain responses reveal the learning of foreign language phonemes, *Psychophysiology* 36 (5), 638–642.

9 Use of keystroke logging to collect cognitive validity evidence for integrated writing tests

Sathena Chan
Centre for Research in English Language Learning and Assessment, University of Bedfordshire, UK

Abstract

Integrated writing tasks are commonly used for teaching, learning and assessment purposes in most higher education contexts. These tasks are cognitively demanding as they require students to transform knowledge by engaging in processes of discourse synthesis, i.e. selecting, organising, and connecting information from multiple source texts into a new or synthesis text. The purpose of the present exploratory study was to investigate L2 writers' discourse synthesis processes underlying the performance of an integrated reading-writing task. The participants were three university students who completed an integrated reading-writing task as part of a post-admission academic literacy test at a British university. Data were collected using a variety of qualitative research techniques: analysis of keystroke logs, retrospective interviews, and text quality analysis. Data analysis revealed distinct engagement in discourse synthesis processes among L2 writers. The study proposes a qualitative approach to analysing keystroke logging data to collect cognitive validity evidence (i.e. test-takers' engagement in discourse synthesis) underlying integrated writing test performance. The other major implications of the findings are the need for explicit teaching and assessment of these discourse synthesis processes, i.e. selecting, connecting and organising relevant ideas from multiple reading stimuli to produce a text, and the need to construct specific rating descriptors which reflect skills of discourse synthesis for integrated writing tasks.

Introduction

There has been a marked increase in language testing research involving keystroke logging, a technology which provides an unobtrusive record of the moment-by-moment composition of a text. It promises to serve as

a powerful tool to collect cognitive validity evidence for writing tests. Cognitive validity concerns 'how closely a writing task represents the cognitive processing involved in writing contexts beyond the test itself' (Shaw and Weir 2007:34). To interpret keystroke logging data as cognitive validity evidence, one major challenge is to relate keystroke logs to models of writing. An additional challenge of collecting cognitive validity evidence for integrated tasks is to capture the intertwined higher-level reading-writing processes as test-takers compose from multiple reading stimuli. The study in this chapter, in response, aims to contribute to this area of research by proposing a methodology that infers the relationships between input, test-takers' discourse synthesis processes and writing output by drawing evidence from these aspects.

Integrated writing tasks are commonly used for teaching, learning and assessment purposes in most higher education contexts. These tasks are cognitively demanding as they require students to transform knowledge by discourse synthesis, i.e. selecting, organising, and connecting information from multiple source texts into a new or synthesis text. The exploratory study examined three university students' engagement in discourse synthesis while completing an integrated test task. The insights that emerged from this new approach to analysing keystroke logging data have important implications for constructing rating descriptors for integrated writing tasks and the incorporation of this tool in assessment.

Literature review

Use of integrated tasks in assessing academic writing

Recently, there has been a renewed interest in using integrated tasks to assess test-takers' academic writing ability. These tasks typically require students to produce a written text in response to one or multiple reading and/or listening stimuli. It is widely agreed that 'the primary rationale for writing tasks that require the integration of content from source material is that, fundamentally, this is what writing for academic purposes involves' (Cumming 2013:3). Alongside independent tasks, integrated tasks can be seen in major English language tests used for making high-stakes decisions about academic admissions. For example, a reading-listening-writing task (with a short passage of about 250 words and a short lecture stimulus) and an independent writing task are used in the Internet-Based Test of English as a Foreign Language (TOEFL iBT). A range of independent and integrated writing tasks including reading-writing (a short passage of up to 300 words) and listening-writing (a short lecture stimulus of 60–90 seconds) tasks are used in the Pearson Test of English (PTE) Academic. A reading-writing task (based on four texts including an infographic of 700 words in total from a

previous reading section) and an independent writing task are used in Trinity College London's Integrated Skills in English (ISE) C1 level. Integrated writing tasks can also be found in regional tests, such as the General English Proficiency Test (GEPT) Advanced test in Taiwan and the Test of English for Academic Purposes (TEAP) in Japan, as well as post-entry academic writing tests in universities (for example see Chan and Latimer 2020, Plakans 2008, Weigle 2002).

With the widespread growth in the popularity of integrated tasks in academic writing tests, studies have investigated the relationships between text quality features of integrated performance and L2 proficiency levels or test scores (e.g. Gebril and Plakans 2016, Knoch, Macqueen and O'Hagan 2014, Plakans and Gebril 2013, Shin and Ewert 2015). Another dimension of integrated writing research is on scoring, for example, to understand the process employed by raters of integrated responses (e.g. Cumming, Kantor and Powers 2001, Gebril and Plakans 2014). Previous studies have also compared students' processes between independent and integrated tasks (e.g. Asención Delaney 2008, Plakans 2008, 2010). Nevertheless, the number of studies that examined test-takers' integrated writing processes remains small (Weir, Vidaković and Galaczi 2013). There is thus a need to gather more empirical evidence to answer fundamental questions concerning the cognitive validity of integrated writing tests. For example, evidence is needed to show 1) whether integrated writing tasks entail the full range of cognitive processes typically required in reading/listening/speaking-writing activities in the academic world beyond the test, 2) whether higher-scoring and lower-scoring test-takers employ these cognitive processes in a different way, and 3) the impact of integrated task features such as modality or the combination/order of stimuli presented on test-takers' processes and performances. Such evidence would be essential to justify the inferences that support the intended interpretations and uses of test scores.

Models of writing and integrated writing processes

In this section I will, firstly, review general models of writing, and then discuss the main cognitive operations involved in discourse synthesis in integrated reading-writing tasks. Hayes and Flower's (1983) model proposes that L1 writing entails three major processes, namely, planning, translating (i.e. translating ideas into linguistic form), and revising. The model is widely used as the theoretical basis for many L2 independent writing studies by means of concurrent think-aloud protocols or retrospective stimulated recalls. It has also been used in studies which investigated the impact of task type (i.e. independent and integrated) on L2 writers' writing. In general, integrated tasks elicited a more recursive process (i.e. writers moving between different processes) for some L2 writers than independent

tasks (Chan 2011, Plakans 2008). Scardamalia and Bereiter's (1987) model explicitly differentiates between two approaches to L1 writing. One is what they call knowledge telling, in which writers construct a text from their own background knowledge on the topic or by retelling appropriate ideas from source texts. The other is knowledge transforming, where the writer transforms knowledge by connecting ideas obtained from source texts and establishing new links between these ideas, or by creating new arguments. The knowledge telling approach to writing focuses on translating pre-existing ideas into a coherent text whereas the knowledge transforming approach tends to involve recursive processes of planning, translating and revising to create new ideas and connections. Most integrated writing tasks require writers to engage in knowledge transforming (Chan 2018, Knoch and Sitajalabhorn 2013).

Kellogg's (1996) is a well-established model of L2 writing, which seems to be the preferred model for studies using keystroke logging (e.g. Kormos 2011, Révész, Michel and Lee 2017). In their review of methodological advances in investigating L2 writing processes, Révész and Michel (2019) explain that Kellogg's model 'puts greater emphasis on the linguistic encoding processes involved in transforming the writer's intended content into text' (2019:493). These explicit accounts help to interpret temporal measures of L2 writing in relation to encoding and execution. To illustrate, if a writer displays frequent long pauses within single words, this most likely indicates the writer's difficulty in lexical retrieval.

In order to form stronger links between cognitive theory and assessing writing practice, Shaw and Weir (2007) proposed a model of L2 writing with a specific focus on six cognitive processes (macro-planning, organisation, micro-planning, translation, monitoring, and revising) which they considered most useful to tell apart skilled and unskilled writers. Their model has influenced many language testing studies over the past decade (e.g. Barkaoui 2016, Bridges 2010). While it has been used extensively in test validation and development projects, its application is largely limited to independent writing. The omission of integrated writing processes in Shaw and Weir's (2007) model, may well reflect a paradigm shift in the assessment of writing from the predominant use of independent tasks in the past when these models were constructed to the current practice of incorporating integrated tasks in language tests.

Writing from sources almost always involves summarisation. According to Hidi and Anderson (1986), summarisation 'requires the comprehension, evaluation, condensation, and frequent transformation of ideas that have been presented' (1986:473). Spivey (1990, 1997) made a distinction between summarising a single text and discourse synthesis from multiple texts. Spivey and King's definition of discourse synthesis has guided this line of research for three decades.

> … discourse synthesis [is] a process in which readers (writers) read multiple texts on a topic and synthesise them. They select content from the composite offered by the sources – content that varies in its importance. They organise the content, often having to supply a new organisational structure. And they connect it by providing links between related ideas that may have been drawn from multiple sources (1989:11).

In her seminal work, Spivey identified three major operations, namely, selection, connection and organisation, involved in discourse synthesis. It has been argued that discourse synthesis is more cognitively demanding than summarising because it requires students to construct their own propositions from multiple sources which present different and sometimes contradictory propositions, and to organise these in a previously non-existent conceptual structure according to the goal of writing (Plakans 2009, Segev-Miller 2007). Chan (2018) extended Shaw and Weir's (2007) model to include processes involved in integrated writing. The study investigated 150 university students' engagement of processes in four integrated writing tasks, two integrated reading-writing tasks under test conditions and two real-world academic writing tasks. Six integrated processes (Table 1) were identified by factor analysis.

Although the model presents the processes in a linear sequence, it is important to note that writing is not conceptualised as a linear sequence of processes, but that the processes often overlap and recur as writing proceeds. For the purpose of this chapter, I will focus on the discourse

Table 1 Cognitive model for integrated writing (Chan 2018)

Processes	Sub-processes
Conceptualisation	Task representation
	Macro-planning
Meaning construction	Careful reading
	Expeditious reading (e.g. skimming, scanning, etc.)
	Selecting relevant ideas from sources
	Connecting ideas
Micro-planning	Micro-planning
Translation	Converting ideas to linguistic form
Organising	Organising ideas in a (new) structure according to the goals of writing
Monitoring and revising	Monitoring and revising during text production (at lower and higher levels)
	Monitoring and revising after text production (at lower and higher levels)

Note: Discourse synthesis processes are in italics.

synthesis processes involved in *meaning construction* and *organising* as they play a much more important role in integrated writing when source materials are more relied on in idea development compared to independent writing.

During meaning construction, the writer constructs meaning based on contextual clues and stimuli (e.g. reading passages) provided in the writing task and their own schematic resources (e.g. background knowledge) (Brown and Yule 1983, Field 2004). When a writer composes from reading sources, various reading comprehension processes are likely to be involved. Khalifa and Weir (2009) distinguish between two types of reading: careful and expeditious reading. Careful reading involves comprehension of every part of the text as a whole. Expeditious reading, on the other hand, involves scanning or skimming the text(s) to access desired information[1]. Most integrated writing tasks, however, require students to go beyond the level of mere reading comprehension. Instead, students are required to engage in discourse synthesis to transform the ideas they have comprehended from the stimuli into their own writing. As argued by Spivey (1990, 1997), writers often need to establish new representations of meaning which integrate source-based propositions and the writer's own knowledge. For example, the meaning representation can be inferences about missing details in the reading source (Kintsch 1974) or the creation of connections between source-based texts and ideas stored in the writer's mind (Seifert, Robertson and Black 1985). In other words, during meaning construction, skilled writers tend to select ideas from sources which are relevant to the writing task and connect these ideas based on their own knowledge of the topic. The process of selecting and connecting ideas during reading-writing activities is, however, an under-researched area because of the complexities in studying real-time interaction between reading and writing. Several recent studies have used keystroke logging with eye tracking to examine reading during writing (e.g. Révész et al. 2017, Van Waes, Leijten and Quinlan 2010), but the component of reading the task or the source material was not within the remit of these studies.

Organising is a process where the writer 'provisionally organises the ideas, still in abstract form, in relation to the text as a whole and in relation to each other' (Field 2004:329). Shaw and Weir (2007) explained that the purpose of organising ideas is to 'determine which are central to the goals of the text and which are of secondary importance' (2007:38). Spivey (1990, 1997) pointed

1 Khalifa and Weir (2009) further argued that depending on the goals of the reader, expeditious reading includes skimming (to obtain the overall gist), scanning (locating predetermined information distributed in a text) and search reading (locating predetermined details at the level of clauses and/or sentences). The reader is also referred to Clarke and Silberstein (1977) for a thorough discussion of different types of reading.

to the additional cognitive demand of organising in discourse synthesis. In integrating and transforming the information that they glean from sources, writers are expected to re-order the ideas, creating a new structure for their own writing that differs from the structure of the sources. Knoch and Sitajalabhorn (2013) argued that the organisation of transformed ideas should be one of the key criteria in assessing integrated writing skills.

I have so far reviewed models of writing and described the key processes involved integrated writing (see Table 1). This study focused specifically on how L2 university students engage in discourse synthesis (Spivey 1990), i.e. selecting, connecting and organising, while completing an integrated writing test task. In the next section, methods which have been used to investigate writers' processes are reviewed.

Methodological approaches for collecting cognitive validity evidence

An understanding of the processes that underpin skilled writing can provide test developers with a systemic view of what a learner is likely to be capable of at different levels of proficiency (Field 2020). Here I briefly review the methods commonly used to collect cognitive validity evidence for writing tests. It is worth noting that most of these studies focus on independent rather than integrated writing tasks.

Methods such as document analysis of a collection of sample writing tasks (e.g. Moore and Morton 2005) and analysis of task demands by expert judgement (Shaw and Weir 2007) are used in some earlier studies. However, these indirect methods have clear limitations, as the anticipated test-takers' processes might not match the actual processes that test-takers use. Questionnaires have been widely used to investigate test-takers' processes in studies focusing on standardised writing tests. For example, Chan, Bax and Weir (2018) and Weir, O'Sullivan, Yan and Bax (2007) conducted larger-scale studies to examine the impact of delivery mode on test-takers' processes on the IELTS independent essay task. Although this method makes it feasible to collect data on a larger scale, questionnaire data only reveal cognitive processes based on the test-takers' recollections. Concurrent verbal reports (Ericsson and Simon 1980), where test-takers provide an account of their cognitive processes while writing, are occasionally used in smaller-scale studies. This is not a particularly popular method for language testing because the process of verbalisation is likely to interfere with the writing processes under investigation (Stratman and Hamp-Lyons 1994). Think-aloud protocols provide data through which researchers can, to some extent, reconstruct the actual cognitive processes, but do not provide a fine-grained analysis *per se*. Retrospective stimulated recall protocols, on the other hand, are widely used to investigate L2 writing processes.

By showing participants a stimulation (e.g. a video of their writing session), this technique intends to reconstruct the processes that test-takers have gone through during writing (Gass and Mackey 2016). However, both concurrent or retrospective verbalisation rely heavily on participants' meta-awareness and their ability to describe the writing processes. Furthermore, as previously mentioned, most of the cognitive validity studies investigated independent rather than integrated writing tasks. Few investigated cognitive processes in integrated writing (especially discourse synthesis) under test conditions. For example, Chan (2018) and Esmaeili (2002) used questionnaires whereas Asención Delaney (2008) and Plakans (2008, 2010) used think-aloud recalls. These studies conclusively point to the challenges of investigating test-takers' engagement in discourse synthesis due to the intertwined nature of the processes. Furthermore, these methods have revealed very little about the interaction between source texts, discourse synthesis processes and outputs.

Recently, keystroke logging, which is a non-obstructive method, has been brought into language testing research, usually supplemented by one or more of the methods discussed above. Keystroke logging technologies record keystroke presses, deletions, pauses and cursor movements, and store the information electronically as log-files for later analysis. This offers a powerful alternative for researchers to record the composition of a text moment-by-moment (Sullivan and Lindgren (Eds) 2006, Van Waes, Leijten, Wengelin and Lindgren 2012). A handful of studies have used keystroke logging to examine the effects of proficiency (Barkaoui 2016, Révész et al 2017, Révész and Michel 2019), task type (Barkaoui 2016, Eklundh and Kollberg 2003) and typing speed (Barkaoui 2016); and to compare L1 and L2 test-takers' revising processes (Chan and Lam forthcoming). Again, most of these studies investigated independent rather than integrated tasks. Some of these studies are reviewed in more depth in the next section.

Previous keystroke logging studies on L2 writing

Révész et al (2017) conducted the most fine-grained keystroke logging study to date on test-takers' processes on an independent writing task. Using keystroke logging and stimulated recall, the study measured temporal features of 30 Chinese test-takers' writing processes on an independent essay task (i.e. IELTS Task 2). The results suggested a positive link between text quality and a number of temporal measures of writing fluency, pausing behaviours and revision behaviours. In general, higher-scoring test-takers (i.e. those who wrote with greater lexical, syntactic and discourse complexity) tended to produce more words per minute, revisit the instructions more and revise more often at higher level than lower-scoring test-takers. On the other hand, compared to higher-scoring test-takers, lower-scoring test-takers

tended to pause within words, to return to a previous word/expression within a paragraph and to look away from the screen during pauses. This shows that fluency measures provided by keystroke logging are particularly useful in revealing writers' automaticity in lexical encoding and formulating sentences with independent tasks. Generally speaking, higher-proficiency writers display higher fluency in translating ideas into linguistic forms, as reflected by temporal measures such as number of words produced by per minute.

In addition to fluency measures, writers' pausing behaviours have been investigated in keystroke logging studies. Frequency, length and location of pauses identified by keystroke logging may indicate some writing processes. By triangulating pause data from keystroke logging with verbal protocol data, researchers have begun to infer which processes are likely to be associated with different types of pauses. It is suggested that pauses between higher textual boundaries of sentences and paragraphs are more likely associated with higher-order processes (e.g., macro-planning), whereas pauses between lower textual boundaries within and between words tend to be linked to lower-level processes such as lexical encoding (Révész et al 2017, Wengelin 2006). There are very few studies on writers' pausing behaviours with integrated tasks. One good example is Spelman Miller's (2000) study which investigated 21 students' pausing behaviours on two academic essay tasks (one descriptive and the other evaluative which requires a synthesis of reading materials). Focusing on the discourse level, the study examined whether and how the students paused at locations of various 'framing devices' (e.g. *it has been stated that*) while composing the two tasks. While the independent and integrated tasks serve rhetorical purposes, the results show that the students composed the tasks in a very similar way in terms of their pausing behaviours. Spelman Miller argued that 'students may fail to identify or respond to the rhetorical demands of assignments in a way expected by teachers and examiners' (2000:142). This points to the need for more research in this area.

Other keystroke logging studies investigated the impact of task type on L2 writers' revision patterns. Eklundh and Kollberg (2003) investigated 10 third year students' revising processes between independent and integrated tasks. The results revealed that most students found the integrated tasks more demanding as they needed to integrate ideas from reading stimuli. This higher demand faced by the students was reflected by higher counts of revision as well as longer pause times and lower text production rates (measured in counts of words in the final texts per minute of writing time). Similarly, Barkaoui (2016) reported that higher proficiency test-takers made slightly more revisions with the integrated task than they did on the independent task, although these differences were not statistically significant. The results also indicate that task type had an impact on when test-takers revised. It was found that while the independent task resulted in a more uniform distribution

of revisions across three stages of writing (i.e. beginning, middle and end), test-takers tended to make more revisions during the middle stage with the integrated task.

From a cognitive validity perspective, to avoid under-representation of the target writing construct, an integrated writing task should elicit responses from test-takers that engage most of the processes proposed in Table 1. It is also important to define what it is that characterises the processing undertaken by a skilled academic writer and what distinguishes it from the processing of a less able academic writer. In addition, as I previously discussed, there should be advances in writing assessment research that develop assessment criteria based on writing processes, not just the quality of the writing product. As demonstrated by the above studies, fluency measures, and pauses and revisions identified by keystroke logging, when interpreted together with verbal protocols, have great potential in revealing some writing processes including planning, formulating and revising. However, it seems that we have not yet fully utilised different types of keystroke logging outputs in writing research.

In addition to temporal measures (the focus of most previous keystroke logging studies), keystroke logging software programs can provide an unobtrusive record of the moment-by-moment creation of the text as well as a record of document switch (e.g. between a particular stimulus and writing sheet) during a reading-writing session. These outputs can potentially provide useful insights into how writers engage in key processes of discourse synthesis, i.e. selecting, connecting and organising, within integrated writing tasks. The present study, in response, aims to contribute to this area of research by proposing a qualitative methodology that enables researchers to make a better alignment between source texts, test-takers' engagement in discourse synthesis and outputs. In particular, the study addresses the following research question (RQ): How do L2 university students engage in discourse synthesis processes when completing an integrated reading-writing task under test conditions?

Methodology

Participants

Participants were three L2 university students at different educational and English proficiency levels at a British university. A pre-sessional student, an undergraduate and a PhD student at different levels of English proficiency participated in the study (see Table 2). Their ages ranged between 18 and 25. They had not been previously assessed on integrated writing tasks. However, being a PhD student, Kelly had more real-world experiences of composing from sources compared to the other two. As the test used in the study was

Table 2 Participants' information

	Educational level	English proficiency level	Discipline	First language
Kelly (F)	PhD	IELTS 7.5[2]	Marketing	Russian
Jacob (M)	Undergraduate	IELTS 6.5	Biomedical science	Croatian
Edward (M)	Pre-sessional	No formal test taken (hence foundation course)	Computer science	Lithuanian

Note: Their names have been changed.

a general academic writing test for all new students (i.e. specific subject knowledge was not required to complete the task), the discipline and first language of the students were not controlled.

Each student completed the task under test conditions, i.e. supervised and timed, for research purposes. Feedback on their performance on the task was given to the students after the test event. However, the test had no stakes for these participants.

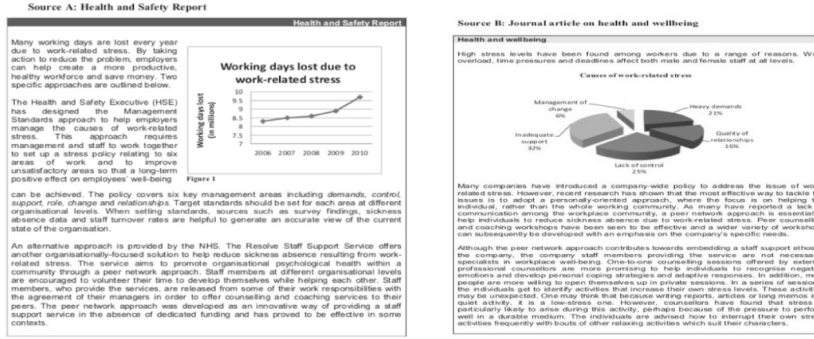
Test task

The reading-into-writing task used in this study is part of an academic literacy test developed at a British university. The test (which includes a reading paper and a reading-into-writing paper) is to be taken by new students after entry to the university. The test aims to identify students who need remedial instruction on academic reading and writing. Based on their test scores, the students are provided with different academic writing support (if needed).

The task requires students to write an essay of at least 250 words in 60 minutes, summarising the main propositions from two one-page passages (each with a non-verbal input, e.g. a chart, a table or a diagram) and providing recommendations on the issue (see Figure 1). The two passages are on the same topic, e.g. work-related stress. Both describe the issue and suggest several solutions to reduce stress in the workplace. The test is designed in such a way that the two passages share a few propositions, e.g. one solution is mentioned in both articles. In this example, students are expected to summarise the issue and three solutions (i.e. top management approach, private one-to-one counselling and peer support approach).

2 Kelly and Jacob took the IELTS test within 12 months of the time of data collection of the study.

Figure 1 An example of the reading stimuli (layout)



At the time of the study, students received a test booklet (which includes the task instruction and the reading stimuli) in hard copy but composed the essay on a computer in a Word document under test conditions. Since then, a fully computer-delivered version of the test was developed and will be implemented shortly.

In this study, a task interface which includes four separate elements, Task Instructions (PDF), Passage 1 (PDF), Passage 2 (PDF) and Writing Sheet (Word document), was used. Participants were allowed to move between elements by clicking the tab of each element in any order. A demonstration of how to navigate the interface was given before data collection and participants were encouraged to have some practice.

Marking

The test adopts an analytic marking scheme whereby the total score is accumulated from scores in three categories which represent the key construct of academic reading-writing skills:

- relevance and adequacy of content (coverage of key points and the transformation of ideas)
- organisation (cohesion and coherence of ideas integrated from the sources)
- language (choice and control of lexis and grammar, the transformation of language).

Examiners can choose from three bands (either 3, 2 or 1) for each criterion. The test has recently been trialled with over 500 students including pre-sessional students, undergraduates and postgraduates. The scale has achieved high levels of rater reliability. Using multi-faceted Rasch measurement, a

further validation study will be conducted to examine the performance of the scale and raters. Based on students' total scores, they will be categorised into three groups: *needs comprehensive support*/*needs some support*/*needs no support* in academic writing skills. Students who are in the first group are recommended (although this is not compulsory) to take an intensive course on academic writing provided by the university. Students who are in the second group are seen by an academic literacy advisor to discuss a self-learning plan to improve their academic writing skills. Students who are in the 'needs no support' category are also welcome to make an appointment with an academic literacy advisor to discuss any questions they may have on academic writing. In this study, participants' essays were examined by a trained rater of the test. The researcher double marked the essays. Their agreement rate was 100%.

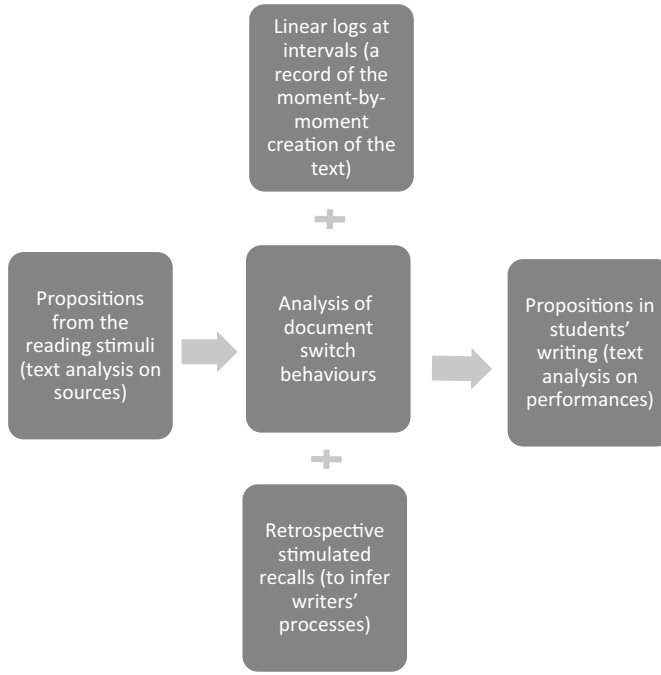
Data collection

Data were collected from one student at a time. After providing written consent to take part in the study, the participant completed a brief background questionnaire. After that, the previously mentioned demonstration of the task interface was given. The participant then composed the integrated writing task on a computer. The time limit for operational test events is 60 minutes. In the study, participants were allowed to finish the sentence they were composing when the time was up. The production of their writing was recorded with the keystroke logging software *Inputlog v7.0* (Leijten and Van Waes 2013). Immediately after completion of the writing task, the participants took part in a stimulated recall session in which they described their composing processes on the task while watching a video recording of their text production. Participants were encouraged to pause the video at any time they wished to describe the thoughts they had at any particular point during the test. The stimulated recall sessions were conducted in English.

Data analysis

In a change from previous keystroke logging studies of L2 writing, the current study employed a qualitative approach (see Figure 2) to investigate test-takers' discourse synthesis processes from the following data sets: propositions from the reading stimuli, propositions in students' writing, analysis of writers' document switch behaviours, linear keystroke logs at intervals, and stimulated recalls. Specifically, evidence of writers' document switch behaviours and chronological keystroke logs will reveal how the participants engage in discourse synthesis in terms of observable behaviours, i.e. selecting, connecting and organising relevant ideas from multiple reading

Figure 2 Data analysis



stimuli to produce their own text. Data from retrospective stimulated recalls will reveal the participants' cognitive processes (see Table 1) underlying these behaviours.

Propositions from the reading stimuli

A checklist of propositions from the two reading stimuli was produced. In this study, propositions refer to individual idea units which can be derived directly from a reading text. A short sentence usually contains a single proposition whereas a compound or complex sentence tends to contain several propositions. To analyse the propositions in the reading stimuli, each passage was divided into T-units, i.e. the smallest possible unit of a sentence that can stand alone grammatically (Hunt 1965). The proposition in each T-unit and the two diagrams was scored for importance using the following scheme (adapted from the aforementioned Plakans and Gebril (2013) work on source text use) by the researcher and double coded by an independent rater:

4: very important (key propositions, i.e. those that should be summarised in the essay)
3: important (supporting propositions)

2: less important (specific details or examples)

1: not important (propositions which are not relevant to the writing task)

A total of 27 propositions were rated as *key propositions*. Table 3 shows the distribution and content of these propositions in the reading stimuli.

Table 3 Distribution and content of key propositions in the reading stimuli

	Passage 1	Passage 2	Diagram 1	Diagram 2	Total
The problem and its causes	1	1	2	3	7
Solution 1 (top management approach)	7				7
Solution 2 (peer support approach)	5 (1 overlapping)	4 (1 overlapping)			8
Solution 3 (one-to-one counselling)		5			5

Propositions in students' writing

The same method was applied to analyse the propositions in the students' writing. Each essay was divided into T-units and each proposition in the T-units was given an importance score. The agreement between raters was above 94% for all propositions in reading stimuli and students' essays.

Evidence from keystroke logging

- *Analysis of writers' document switch behaviours*

The task interface includes four documents, i.e. Task Instructions, Passage 1, Passage 2 and Writing Sheet. The pattern of how the writers switched between these documents was analysed in terms of frequency of switches, order of switches and length of switches.

- *Linear keystroke logs at intervals*

A chronological basic log file, which provides a linear step by step representation of the production of a text, was produced for each writer. In this study, the log outputs were produced at focus-based intervals with each document switch set as the focus.

Evidence from stimulated recalls

Writers' retrospective recalls of their writing processes were coded into process categories based on the model presented in Table 1. An independent rater double-coded all categories and the agreement was 96%. In this chapter, we will only discuss the results in relation to five sub-processes: careful reading, expeditious reading, selecting ideas, connecting ideas and organising. Examples of coding are presented in Table 4.

Table 4 Examples of stimulated recall protocols

Processes	Examples
Careful reading	• I read every sentence carefully and made notes.
Expeditious reading	• I tried to get an overall gist, the topic of the two articles first.
Selecting ideas	• I didn't include all the details of this [Solution 1] because I really couldn't relate to it. I didn't get it. I mean I can understand what it means but how this would help to reduce work-related stress.
Connecting ideas	• I just realised that peer network was another solution. • I knew this [Solution 2] was also mentioned in the first passage. So I need to combine the information from both sources.
Organising	• I was moving the sentences to make it more coherent. I tried to use connectives to link the ideas, to show the relationship.

Results

To remind the reader, the study investigated how L2 university students engage in discourse synthesis processes when completing an integrated reading-writing task under test conditions. Students' performance on the test will be reported, followed by findings regarding their engagement in discourse synthesis.

Students' performance: scores and summarisation of propositions

As described in the methodology section, students can score 1, 2 or 3 on each criterion, totalling a maximum score of 9. Kelly and Jacob scored 8 out of 9 and Edward 5; see Table 5.

Table 5 Students' test scores

Scoring criteria	Kelly	Jacob	Edward
Relevance and adequacy of content	2	2	1
Compositional organisation	3	3	2
Language	3	3	2
Total task score	8	8	5

As the focus of the study was to investigate students' discourse synthesis processes, their scores on *relevance and adequacy of content* are further discussed here. Edward's ability to summarise was weakest among the three, hence rated 1 out of 3 on this criterion. Although he managed to mention all three solutions in his essay, he omitted most of the propositions within each solution. Both Kelly and Jacob had the same score (2 out of 3) but they included different propositions. For example, Kelly summarised

almost all of the key propositions of the three solutions, drawing from both passages, but she hardly made use of information from the two diagrams. Jacob, on the other hand, included propositions from all sources (i.e. two articles and two diagrams) but he failed to summarise Solution 3. The differences in their summarisation processes will be discussed in more depth later.

The next section reports how the students summarised propositions of Solution 2. This is chosen as an example for illustration because the students need to recognise that Solution 2 is in fact mentioned in both passages.

Discourse synthesis processes on the integrated writing task

This section reports how the three students summarised Solution 2 (i.e. the peer support approach) in their writing (summarise and discourse synthesis are used interchangeably hereafter). To remind the reader, there are eight propositions relevant to Solution 2 (see Table 3). Kelly included six of the propositions (four from Passage 1 and three from Passage 2; two overlapping propositions were counted as one), Jacob included four (three from Passage 1 and one from Passage 2) and Edward two (two from Passage 1 and none from Passage 2). Tables 6, 7 and 8 present chronological segments of the students' text production in relation to Solution 2.

Edward (careful reading and note taking at the beginning; limited reading during summarisation; some local expeditious reading)

As mentioned previously, among the three students, Edward was least successful in summarising Solution 2. He included only two of the eight propositions and he failed to include any relevant ideas from Passage 2. At the beginning of his task time, he spent almost 17 minutes reading through the two passages carefully and taking notes before he started to write the essay (00:16:48). He explained in the interview that he 'read every sentence carefully and made notes'. While reading Passage 1, he made the following note (square brackets indicate observation notes added by the researcher):

> Peer network *[Solution 2]*: collect survey findings etc to see accurate view of situation *[Solution 1] [At present the two solutions are muddled]*

Edward's notes suggested that his comprehension was not fully accurate as he did not distinguish between the two solutions in Passage 1. When Edward summarised, he relied mostly on his notes and seldom returned to the passages.

Table 6 shows the sequential linear output of a segment of his summarising processes (0:22:27–0:24:07). At this point, he had finished summarising Solution 1 (i.e. to set up a policy). At 0:22:27, he returned to read Passage

1 again. He explained that he 'wasn't sure how peer network fits into the management approach [solution 1]'. After reading, he realised that 'peer network was another solution'. At 0:22:39, he introduced Solution 2 in his writing. He returned to Passage 1 at 0:24:02 but only stayed for two seconds. He then moved on to summarise Solution 3.

In short, Edward's summarisation of Solution 2 was shallow, probably because of the gap in his comprehension during meaning construction. During this period, he had three document switches, all between Passage 1 and the Writing Sheet. Although Edward was able to identify the gap in his notes during writing, he did not pick up any other propositions about Solution 2 from Passage 1, nor did he realise Solution 2 was also mentioned in Passage 2. He only introduced Solution 2 but failed to summarise most of the relevant propositions.

Table 6 Segments of Edward's text production (0:22:27–0:24:07)

Document	Starting time	Duration	Liner keystroke logs
Passage 1	0:22.27	0:00:12	
Writing Sheet	0:22:39	0:01:23	In add###### Furthermore, another reasen######### solution was given,# peer network
Passage 1	0:24:02	0:00:05	
Writing Sheet	0:24:07	0:00:39	approach, where employers could spend their work time for helping each others and also helping themselfs.

Note: # indicates a deletion of the previous character.

Jacob (expeditious reading at the beginning; careful reading of relevant parts of one passage during summarisation; added propositions from another passage during revision)

Jacob's summarisation of Solution 2 consisted of two major segments. He summarised five of the eight propositions about Solution 2 but he drew more heavily from Passage 1. Jacob began the task by reading the task instructions and the two passages for about seven minutes. He did not take any notes. At 0:07:25, he started to compose his first paragraph to describe the issue and its causes. He then composed his second paragraph about Solution 1. Table 7 shows the linear segments of his text production in relation to Solution 2. At this point (0:27:55), he had just finished writing about Solution 1 and turned to Passage 1 for about nine seconds. Some interesting observations about Jacob's summarising process can be made. First, he referred to Passage 1 regularly when he wrote about Solution 2 between 0:27:55 and 0:36:46. During this segment, Jacob had frequent switches (14 in total) between Passage 1 and the Writing Sheet, but he did not refer to Passage 2, nor did he

include any propositions from Passage 2. After this, he moved on to evaluate the two solutions and write his recommendations. Towards the end of the task time at 0:52:56, Jacob turned to Passage 2 and he added a proposition about the weakness of Solution 2 in his concluding paragraph. Later, he moved this sentence (*One of the problems that may arise is that volunteers may often not be professionals*) to the third paragraph, where Solution 2 was described.

Table 7 Segments of Jacob's text production (0:27:55–0:36:46 and 0:52:56–0:54:31)

Documents	Starting time	Duration	Keystroke logs
Passage 1	0:27:55	0:00:09	
Writing Sheet	0:28:04	0:02:24	As a second solution, more individually approach was introduced. The main characteristics of this approach was to focus on the individual
Passage 1	0:30:28	0:00:23	
Writing Sheet	0:30:51	0:00:16	characteristic
Passage 1	0:31:07	0:00:44	
Writing Sheet	0:31:51	0:00:02	n/a
Passage 1	0:31:53	0:00:01	
Writing Sheet	0:31:54	0:00:07	rather than
Passage 1	0:32:01	0:00:10	
Writing Sheet	0:32:11	0:00:54	instead of the whole commu########### individual ###########
Passage 1	0:33:05	0:00:29	
Writing Sheet	0:33:34	0:00:01	n/a
Passage 1	0:33:35	0:00:26	
Writing Sheet	0:34:01	0:00:28	, known as
Passage 1	0:34:29	0:00:05	
Writing Sheet	0:34:34	0:02:12	peer-network approach################ ######## Since ###### It was planned that the staff members would provide a service and be######## volunteer organised in a way that staff would volunteer to help others to establish better communication between themselves
Passage 2	0:52:56	0:01:35	
Writing Sheet	0:54:31	0:00:05	One of the problems that may arise is that volunteers may often not be pro### spec#### professionals

Note: # indicates a deletion of the previous character.
Note 2: n/a indicates that the writer did not type anything within the duration.

Kelly (expeditious reading at the beginning; both careful and expeditious reading of relevant parts during discourse synthesis; flexible in switching between reading, discourse synthesis and revising)

Kelly's summarisation of Solution 2 consisted of three major segments and the processes were most recursive among the three students. Kelly included six of the eight propositions by drawing from both Passage 1 and Passage 2. Like Jacob, Kelly began the task by skimming the task instructions and the two passages for about five minutes. At 0:05:02, she started her essay by describing the issue. As she did not make use of the diagrams, she missed most of the propositions about the issue and its causes. She then summarised the propositions of Solution 1 briefly. She explained in the stimulated recall that she struggled to 'relate to' Solution 1. She was able to understand the literal meaning, but she could not see 'how this would help to reduce work-related stress'. It is interesting to note that Kelly made a deliberate choice to not include a particular proposition, a (de)selecting process which was not observed from the other two students.

At 0:14:37, Kelly had finished summarising the propositions about Solution 1. Table 8 shows the three segments of her summarisation of Solution 2 (i.e. 0:14:37–0:22:55, 0:26:20–0:40:25 and 0:43:42–0:48:03). During the first segment (0:14:37–0:22:55), she had five switches between Passage 1 and the Writing Sheet. Her summarisation process was intertwined with selective reading of Passage 1. Minimal revisions were made when she summarised from a single source (i.e. Passage 1) during the first segment.

At the beginning of the second segment (0:26:20–0:40:25), she spent almost six minutes re-reading parts of Passages 1 and 2 carefully. As she explained in the interview, when she skimmed the passages at the beginning, she knew that Solution 2 was mentioned in both passages and she needed to 'combine the information from both sources'. This was followed by a writing period of about six minutes, a period when she was trying to integrate propositions from the two passages. During this segment, Kelly made more revisions, including both deletions at the point of inscription and substitutions at a previous point of inscription (indicated by # and <> in Table 8). Kelly then moved on to revise the propositions on Solution 1.

At 0:43:42 (the beginning of the third segment), Kelly resumed to revise the propositions on Solution 2. Again, she made some substitutions at a previous point of inscription, followed by a writing period of almost three minutes. After that, she paused to return to Passage 2 briefly for two seconds before she summarised the disadvantages of Solution 2. She explained in the stimulated recall that she turned to Passage 2 to remind herself of the final idea she planned to include.

Table 8 Segments of Kelly's summarisation of Solution 2 (0:14:37–0:22:55, 0:26:20–0:40:25 and 0:43:42–0:48:03)

Documents	Starting time	Duration	Keystroke logs
Passage 1	0:14:37	0:00:26	
Writing Sheet	0:15:03	0:05:37	Another way of managing stress in organisations is support from the NHS that aims to to encourage a healthy psychological environment through a peer networking. In this case managers need to support such initiatives by giving staff memebers that participate in peer netwroking some priviliges such as reducing
Passage 1	0:20:40	0:00:06	
Writing Sheet	0:20:46	0:00:55	those the level fo responsibilities of work.
Passage 1	0:21:41	0:00:18	
Writing Sheet	0:21:59	0:00:54	Such approach works especially well if a company has no particular funding for initiating stress management programs.
Adding propositions to describe the causes of the issue			
Passage 1 and 2	0:26:20	0:05:51	
Writing Sheet	0:32:11	0:06:23	Alternatively, a company migh aim to employ a ############arget individal employees instead of the whole ######################, which is considered to be more of a a personally informed approach rather than a community-oriented one. The reason for emplying such approach is ########################## ###############hat approqaqch is ################## personally informed approach is more effective as it #. Moreover, ##########, it also helps managres to target individuals in a company and thus improve communication between work peers. such as coaching programs ###### by employing ###, ## or peer counselling as an example #######. These programs ########will help with ############promoting ##is###or## Another way of promoting is This is considered to be one of the most effective way of stress management at ################## ## ############################## ## ############################ ###targets individuals instead of community
Writing Sheet (revisions)	0:38:34	0:00:02	<managres to target individuals in a company and thus improve communication between work peers.>

Table 8 (continued)

Documents	Starting time	Duration	Keystroke logs
Writing sheet	0:38:36	0:01:49	instead of a community helps ######aims to##, which makes this approach especially effective among all others.
Revising propositions on Solution 1			
Writing Sheet (revisions)	0:43:42	0:00:02	<h approach helps managres to target individuals in a company instead of a community and thus aims to improve communication between work peers, which makes this approach especially effective among all others>
Writing Sheet	0:43:44	0:02:54	To employ this approach successfully, ######employees that are willing to take part##########participate in peer netwroking by ###### for example. ##Peer-netwroking approach #
Passage 2	0:46:38	0:00:02	
Writing Sheet	0:46:40	0:01:23	other##########. However, peer-networking also has its disadvanatges. One of them is the lack of employees' comptenece in the area of stress management.

Note: # indicates a deletion of the previous character.
Note 2: <> a chunk of text which was moved to a different part of the essay.

Discussion

Use of keystroke logging to infer higher-level discourse synthesis processes

One advantage of keystroke logging is that it records objective data of students' real-time writing behaviours when composing a task. In the tradition of psycholinguistic research into text production, temporal features are viewed as 'naturally displayed sources of evidence' (Garrett 1982:23) of real-time language processing. The analyses of pauses during writing activities have been used to investigate planning, as measures of pauses provide important insights into writers' allocation of attentional resources to otherwise hidden processes during composition (Kellogg 1994). For integrated writing, these 'hidden' processes can be construct-relevant activities such as planning, meaning construction, and reviewing, or construct-irrelevant activities such as daydreaming. The findings of this exploratory study show evidence of the ways three university students engaged in discourse synthesis processes, i.e. selecting, connecting, and

organising ideas from multiple reading stimuli when composing an integrated reading-writing test task.

Edward, the lowest-scoring student, approached the integrated task largely in two separate stages, reading the source carefully and then summarising the propositions. He relied on the knowledge telling approach (Scardamalia and Bereiter 1987). When he summarised, he mainly drew upon his understanding of the passages. Even when he noticed the gap in his initial plan (i.e. the omission of Solution 2), he was reluctant to read the passages again to summarise Solution 2. As a result, the propositions he summarised were inadequate.

On the other hand, Jacob and Kelly, the high-scoring students, adapted more of the knowledge transforming approach through discourse synthesis processes to complete the task. Both students started with expeditious reading to get the gist of the source. During discourse synthesis, they interacted with the stimuli regularly, though in different patterns. Jacob summarised propositions from a single passage during writing, which indicates that his ability to synthesise from multiple sources was limited. However, he compensated for this during his revising processes. As he revised, he was able to connect the propositions from the two passages by moving sentences around and using cohesive devices, e.g. *however*. Kelly was flexible in her discourse synthesis processes, integrating propositions from both passages. She switched swiftly between careful and expeditious reading, synthesising from multiple stimuli and revising. It is important to note that pauses in her discourse synthesis processes were not held due to lack of fluency in execution (automatisation) but a reflection of meaning construction processes. For example, she reported in the stimulated recall that she was hesitant to summarise propositions which she did not necessarily agree with. This process of de(selecting) propositions was not observed from the other two students.

Alignment between keystroke logging outputs and cognitive processes

Apart from being labour-intensive, the biggest challenge of analysing keystroke logging data would be to build a theory-informed alignment between keystroke logging outputs of text production and the underlying writing processes. This study demonstrates how two types of outputs (i.e. a record of document switch behaviours and linear keystroke logs at intervals) could be interpreted in conjunction with the text analysis of summary propositions and students' retrospective accounts to illuminate students' discourse synthesis under test conditions. The differences in the three students' discourse synthesis processes could be useful indicators of their ability to compose a text from multiple sources, which have long been considered as an important construct of academic writing (Asención

Delaney 2008, Chan 2018, Gebril and Plakans 2016). The approach to analysing keystroke logging data proposed in this chapter would help to illuminate differences in test-takers' processes which would not otherwise appear in the final product. The linear representation of text production generated by keystroke logging reveals the different ways of composing an integrated writing task among students. The function of searching and filtering linear keystroke logs offers potential for researchers to evaluate certain higher-level processes such as building links to connect ideas from multiple stimuli. With previous methods like concurrent think-aloud protocols and retrospective questionnaires, researchers relied heavily on students' self-reporting of their writing processes. Keystroke logging opens up opportunities to gather evidence of test-takers' discourse synthesis from the cognitive perspective of real-time language processing. The question for language testers is how some of the data about test-takers' writing processes could be incorporated into the assessment cycle.

Implications for L2 writing assessments

With the use of keystroke logging, as demonstrated in the current study, researchers are now able to better identify students' engagement in discourse synthesis processes with integrated writing tasks. As mentioned previously, the task used in this study was designed to provide a post-entry assessment to all new students at a British university to indicate whether they need comprehensive, some or no support in academic writing skills. Based on their test scores, Edward (scored 5 out of 9) would have been recommended for comprehensive support and be taught an academic writing course whereas Kelly and Jacob (both scored 8 out of 9) would have been recommended to improve their academic writing through self-learning materials. The test scores have indicated the right student who would need training on discourse synthesis most. On the other hand, it would be important to consider the extent to which current scale descriptors can adequately capture differences in discourse synthesis performance, particularly between Kelly and Jacob. Kelly seems to be the one displaying the highest skill as a writer in terms of engaging with source materials and even de-selecting propositions. This is probably because she is a PhD student and therefore likely to have had more experiences of integrated writing in an academic context compared to the other two. The descriptors were designed to have three bands only because the test requires an efficient marking system to evaluate a large number of performances within a very short window of score reporting. However, the findings of this study suggest that the current rating scale might not necessarily be able to capture nuances in discourse synthesis skills (although it is fit for purpose in identifying candidates who need more support).

One may also argue that the students are not fully benefitting from the test event unless they are provided with learning-oriented feedback to specify what they need to improve. As demonstrated in this study, keystroke logging data help to reveal differences in L2 writers' engagement in discourse synthesis processes, i.e. selecting, connecting and organising ideas from multiple reading stimuli. One important question for language testing is how keystroke logging could in fact feed into a grade or provide learning-oriented feedback.

It appears that in terms of constructing descriptors for integrated writing tasks, students' ability to select ideas from multiple stimuli can be evaluated by a 'content' criterion in relation to the inclusion of main propositions. Students' ability to connect and organise ideas can possibly be evaluated by the 'coherence' of the integrated ideas and 'structure' of the writing. This, however, would require some prompt-specific training as the raters would have to consider not only the structure of the essay (i.e. how different parts of an essay fit together coherently – which is a feature assessed in most writing tests) but also the links between ideas transformed from stimuli. Text mining techniques which can automatically generate relationship labels and a hierarchy of ideas might offer new ways for raters (both human raters and e-raters) to evaluate text features underlying students' discourse synthesis processes. Another implication is that the proposition analysis used in this study can perhaps be simplified and used in descriptors, i.e. if those are identified in reading texts, then raters can use a checklist to determine which ones have been subsequently used in writing.

Conclusions – keystroke logging as evidence for assessment?

In several important ways, the investigation of test-takers' discourse synthesis processes in this study departs from and extends approaches taken in previous keystroke logging studies on L2 writing. The exploratory study, however, has several limitations. The three students were recruited at different educational levels and proficiency levels, but the sample is too small to be representative of the general population of students in these categories. The fact that the participants come from three different disciplines could also have introduced another confounding factor on their writing processes.

To conclude, I reflect on whether process-tracking technologies such as keystroke logging can help language testers to move towards incorporating evidence of writing processes into assessment criteria (which currently consider only features of the writing product). While almost all existing standardised writing tests measure students' writing abilities through their final written products, it is evident that data captured by keystroke logging can provide rich information about the temporal measures of test-takers'

writing and their discourse synthesis processes. Many researchers argue that such data could support formative assessment and tutorial activities such as learner–tutor discussions and learner–learner co-operative writing sessions (Ranalli, Feng and Chukharev-Hudilainen 2018, Sullivan, Kollberg and Palsson 1997). For example, Lindgren and Sullivan's (2003) study demonstrates that keystroke logging can facilitate reflection on the gaps in writing processes (e.g. making higher-level revisions) by replaying the writing session in classrooms. Although most high-stakes writing tests are now either computer-delivered or have a computer-delivered version, we are still quite far away from harnessing process-tracing techniques into online testing systems. For that to happen, we need to overcome the various hurdles of analysing and interpreting keystroke logging data discussed in this chapter. In particular, we need to establish a solid research effort to answer some of the following outstanding questions: What would have to be considered to ensure use of test-takers' processing data in assessment is ethical? What is the predictive power of temporal measures associated with fluency, pauses and revisions on L2 writers' integrated writing skills? To what extent can keystroke logging data or other process-tracking measures inform a grade in the target language use domain? How could keystroke logging in fact feed into a grade or provide learning-oriented feedback?

It is hoped that this chapter has made a methodological contribution towards this research effort by proposing one way of building better alignment between keystroke logging outputs and higher-level integrated writing processes.

References

Asención Delaney, Y (2008) Investigating the reading-to-write construct, *Journal of English for Academic Purposes* 7 (3), 140–150.

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

Bridges, G (2010) Demonstrating cognitive validity of IELTS Academic Writing Task 1, *Research Notes* 42, 24–33.

Brown, G and Yule, G (1983) *Discourse Analysis*, Cambridge: Cambridge University Press.

Chan, S (2011) Demonstrating cognitive validity and face validity of PTE Academic Writing items Summarize Written Text and Write Essay, *Pearson Research Note*, 1–20.

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

Chan, S and Lam, D (forthcoming) *Investigating the textual features and revising processes of EFL and L1 English writers in China*, ARAGs Research Reports Online.

Chan, S and Latimer, N (2020) Placing construct definition at the heart of language assessment: research, design and a priori validation, in Saville, N and Taylor, L (Ed) *Lessons and Legacy: A Tribute to Professor Cyril J Weir (1950–2018)*, Studies in Language Testing Volume 50, Cambridge: UCLES/Cambridge University Press, 105–131.

Chan, S, Bax, S and Weir, C J (2018) Researching the comparability of paper-based and computer-based delivery in a high-stakes writing test, *Assessing Writing* 36, 32–48.

Clarke, M A and Silberstein, S (1977) Toward a realization of psycholinguistic principles in the ESL reading class, *Language Learning* 27 (1), 135–154.

Cumming, A (2013) Assessing integrated writing tasks for academic purposes: Promises and perils, *Language Assessment Quarterly* 10 (1), 1–8.

Cumming, A H, Kantor, R and Powers, D E (2001) *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: an investigation into raters' decision making and development of a preliminary analytic framework*, TOEFL Monograph No. MS-22, Princeton: Educational Testing Service.

Eklundh, K S and Kollberg, P (2003) Emerging discourse structure: computer-assisted episode analysis as a window to global revision in university students' writing, *Journal of Pragmatics* 35 (6), 869–891.

Ericsson, K A and Simon, H A (1980) Verbal reports as data, *Psychological Review* 87 (3), 215–251.

Esmaeili, H (2002) Reading-to-write reading and writing tasks and ESL students' reading and writing performance in an English language test, *The Canadian Modern Language Review* 58, 599–622.

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

Field, J (2020) Cyril Weir and cognitive validity, 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, 54–82.

Garrett, M F (1982) Production of speech: observations from normal and pathological language use, in Ellis, A W (Ed) *Normality and Pathology in Cognitive Functions*, London: Academic Press, 19–76.

Gass, S and Mackey, A (2016) *Stimulated Recall in Second Language Research* (Second edition), New York: Routledge.

Gebril, A and Plakans, L (2014) Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks, *Assessing Writing* 21, 56–73.

Gebril, A and Plakans, L (2016) Source-based tasks in academic writing assessment: Lexical diversity, textual borrowing and proficiency, *Journal of English for Academic Purposes* 24, 78–88.

Hayes, J R and Flower, L (1983) Uncovering cognitive processes in writing: An introduction to protocol analysis, in Mosenthal, P, Tamor, L and Walmsley, S (Eds) *Research on Writing: Principles and Methods*, New York: Longman, 207–220.

Hidi, S and Anderson, V (1986) Producing written summaries: Task demands, cognitive operations, and implications for instruction, *Review of Educational Research* 56, 473–493.

Hunt, K (1965) A synopsis of clause-to-sentence length factors, *The English Journal* 54, 305–309.

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

Kellogg, R T (1996) A model of working memory in writing, in Levy, C M and Ransdell, S (Eds) *The Science of Writing: Theories, Methods, Individual Differences and Applications*, Mahwah: Lawrence Erlbaum Associates, 57–71.

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

Kintsch, W (1974) *The Representation of Meaning in Memory*, Hillsdale: Lawrence Erlbaum Associates.

Knoch, U, Macqueen, S and O'Hagan, S (2014) An investigation of the effect of task type on the discourse produced by students at various score levels in the TOEFL iBT® Writing test, *ETS Research Report Series 2014* (2), 1–74.

Knoch, U and Sitajalabhorn, W (2013) A closer look at integrated writing tasks: Towards a more focused definition for assessment purposes, *Assessing Writing* 18 (4), 300–308.

Kormos, J (2011) Task complexity and linguistic and discourse features of narrative writing performance, *Journal of Second Language Writing* 20, 148–161.

Leijten, M and Van Waes, L (2013) Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes, *Written Communication* 30 (3), 358–392.

Lindgren, E and Sullivan, K P H (2003) Stimulated recall as a trigger for increasing noticing and language awareness in the L2 writing classroom: A case study of two young female writers, *Language Awareness* 12, 172–186.

Moore, T and Morton, J (2005) Dimensions of difference: a comparison of university writing and IELTS writing, *Journal of English for Academic Purposes* 4, 43–66.

Plakans, L (2008) Comparing composing processes in writing-only and reading-to-write test tasks, *Assessing Writing* 13, 111–129.

Plakans, L (2009) The role of reading strategies in integrated L2 writing tasks, *Journal of English for Academic Purposes* 8 (4), 1–15.

Plakans, L (2010) Independent vs integrated writing tasks: A comparison of task representation, *TESOL Quarterly* 44 (1), 185–194.

Plakans, L and Gebril, A (2013) Using multiple texts in an integrated writing assessment: Source text use as a predictor of score, *Journal of Second Language Writing* 22 (3), 217–230.

Ranalli, J, Feng H-H and Chukharev-Hudilainen, E (2018) Exploring the potential of process-tracing technologies to support assessment for learning of L2 writing, *Assessing Writing* 36, 77–89.

Révész, A and Michel, M (2019) Methodological advances in investigating L2 writing processes: Introduction, *Studies in Second Language Acquisition* 41 (3), 491–501.

Révész, A, Michel, M and Lee, M (2017) *Investigating IELTS Academic Writing Task 2: Relationships between cognitive writing processes, text quality, and working memory*, IELTS Research Report Online Series.

Scardamalia, M and Bereiter, C (1987) Knowledge telling and knowledge transforming in written composition, in Rosenberg, S (Ed) *Advances in Applied Psycholinguistics Volume 2: Reading, Writing and Language Learning*, Cambridge: Cambridge University Press, 142–175.

Segev-Miller, R (2007) Cognitive processes in discourse synthesis: The case of intertextual processing strategies, in Rijlaarsdam, G, Torrance, M, Van Waes, L and Galbraith, D (Eds) *Writing and Cognition: Research and Applications*, Amsterdam: Elsevier, 231–250.

Seifert, C M, Robertson, S P and Black, J B (1985) Types of inferences generated during reading, *Journal of Memory and Language* 24, 405–422.

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

Shin, S-Y and Ewert, D (2015) What accounts for integrated reading-to-write task scores?, *Language Testing* 32 (2), 259–281.

Spelman Miller, K (2000) Academic writers on-line: investigating pausing in the production of text, *Language Teaching Research* 4 (2), 123–148.

Spivey, N N (1990) Transforming texts: Constructive processes in reading and writing, *Written Communication* 7 (2), 256–287.

Spivey, N N (1997) *The Constructivist Metaphor: Reading, Writing and the Making of Meaning*, San Diego: Academic Press.

Spivey, N N and King, J R (1989) Readers as writers composing from sources, *Reading Research Quarterly* 24 (1), 7–26.

Stratman, J F and Hamp-Lyons, L (1994) Reactivity in concurrent think-aloud protocols, in Smagorinsky, P (Ed) *Speaking about Writing: Reflections on Research Methodology*, Thousand Oaks: Sage, 89– 111.

Sullivan, K P H and Lindgren, E (Eds) (2006) *Computer Keystroke Logging and Writing*, Oxford: Elsevier Science.

Sullivan, K P H, Kollberg, P and Palsson, E (1997) L2 writing: a pilot investigation using keystroke logging, in Diaz, L and Pérez, C (Eds) *Views on the Acquisition and Use of a Second Language*, Barcelona: Universitat Pompeu Fabra, 553–566.

Van Waes, L, Leijten, M and Quinlan, T (2010) *Reading During Sentence Composing and Error Correction: A Multilevel Analysis of the Influences of Task Complexity*, The Netherlands: Springer.

Van Waes, L, Leijten, M, Wengelin, Å and Lindgren, E (2012) Logging tools to study digital writing processes, in Berninger, V W (Eds) *Past, Present, and Future Contributions of Cognitive Writing Research to Cognitive Psychology*, New York: Psychology Press, 507–533.

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

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

Weir, C J, O'Sullivan, B, Yan, J and Bax, S (2007) *Does the computer make a difference? The reaction of candidates to a computer-based versus a traditional hand-written form of the IELTS Writing component: effects and impact*, IELTS Research Reports Volume 7.

Wengelin, Å (2006) Examining pauses in writing: Theory, methods and empirical data, in Sullivan, K P H and Lindgren, E (Eds) (2006) *Computer Keystroke Logging and Writing*, Oxford: Elsevier Science,107–130.

Part 2
Using technology to enhance language assessment

10 Video-conferencing speaking tests: An investigation of context validity related to test administration

Chihiro Inoue
Centre for Research in English Language Learning and Assessment, University of Bedfordshire, UK

Fumiyo Nakatsuhara
Centre for Research in English Language Learning and Assessment, University of Bedfordshire, UK

Vivien Berry
Formerly British Council, UK

Evelina Galaczi
Cambridge University Press & Assessment, UK

Abstract

Face-to-face speaking assessment provides the benefit of eliciting a broad interactional construct, but at the cost of being logistically complex, resource-intensive and difficult to manage. Advances in video-conferencing (VC) technology now make it possible to engage in online interaction more successfully than previously, thus reducing dependence upon physical proximity between the examiner-interlocutor and the candidate(s). It is therefore not surprising that such technology is seen as a valuable assessment tool in geographically remote and politically unstable areas of the world, or indeed in contexts affected by the social distancing required during the recent Covid-19 pandemic. However, the administrative conditions under which the test takes place, one of the key contextual parameters of the VC-delivered test, is often overlooked, despite its potentially significant influence on candidates' performance and therefore overall test validity (Weir 2005).

In this chapter, we report on investigations into administrative features of a VC-delivered high-stakes speaking test, including aspects of examiner behaviour and the effectiveness of examiner training for the VC test delivery. The chapter ends with a discussion of how administrative settings and the management of speaking tests play a key role in ensuring the context validity of VC speaking tests. It also offers suggestions for the operationalisation of a VC IELTS Speaking test, as well as broader implications for test administration and examiner training in other VC tests.

Introduction

Digital technologies are transforming the way people communicate. Video-conferencing (VC) applications such as Zoom, WhatsApp and WeChat are now prevalently used in personal, professional and educational settings. VC has also attracted language testers' attention, largely due to its potential to tap into a construct of speaking which includes interaction, thus addressing the construct under-representation of computer-based speaking tests while reducing the logistically complex resource-intensive demands of face-to-face test administration (Bernstein, Van Moere and Cheng 2010, Galaczi 2010, O'Loughlin 2001, Qian 2009, Xi 2010).

The practical benefits of VC-delivered tests have led to a proliferation of research into the comparability of the two delivery modes. A number of investigations have been carried out, focusing on score comparability (e.g., Clark and Hooshmand 1992, Kim and Craig 2012, Nakatsuhara, Inoue, Berry and Galaczi 2016, 2017a, 2017b), examiner perceptions (Clark and Hooshmand 1992, Nakatsuhara et al 2016, 2017b), and examiner rating behaviours (Nakatsuhara et al 2016, 2017b). These studies provided evidence of test comparability, despite some minor differences in examiner perceptions of the two modes and candidates' use of language functions which some researchers concluded are surmountable through enhancement of examiner and candidate training on the VC tests (Clark and Hooshmand 1992, Nakatsuhara et al 2017b).

An apparent notable gap in the literature on the comparability of face-to-face and VC tests, however, is the lack of reports on how examiners were trained to conduct the VC mode of speaking tests, as well as on how effective such examiner training had been in influencing their exam delivery behaviours. The content of VC examiner training and the evaluation of its effectiveness are crucial because, according to Weir's socio-cognitive framework for speaking test validation (Weir 2005, further elaborated in Taylor 2011), the condition under which a test is administered is an important contextual parameter that can significantly influence candidates' performance and overall test validity (Weir 2005:46). While there is a body of research on the variability of examiner behaviours and its considerable

impact on the context and scoring validity of face-to-face speaking tests (Brown 2003, Lazaraton 2002, Nakatsuhara 2008), very little has been reported on the same factor for VC-delivered tests. Ensuring the effectiveness of VC examiner training warrants comparability of test administration, elicited candidate performance, and subsequently, the scores between the face-to-face and VC modes of a speaking test.

With that research gap in mind, this chapter reports on the content of examiner training designed for the VC mode of the IELTS Speaking test and investigates the effectiveness of the VC examiner training, which provides validity evidence for the VC mode of the test. Despite being a small-scale study involving a total of 18 trained and certified IELTS Speaking Examiners, and thus with limited generalisability (see the section 'The context of the current study'), the training procedures described in this chapter and their proven effectiveness based on examiner feedback provide useful, practical guidance for other VC tests.

Literature review

When considering test administration as a contextual parameter in two test modes which exist in parallel in different geographic contexts, such as the VC and face-to-face modes of the IELTS Speaking test, it is important to investigate test elements such as the examiner script and examiner behaviour across the two modes. A key consideration is whether there are equal opportunities for test-takers to demonstrate their speaking skills in both modes. Such a comparability investigation would provide evidence for the criterion-related validity of the test modes (Weir 2005).

Providing equal opportunities here does not mean using identical test procedures to offer the same test experiences. Although communication via VC is much closer to face-to-face communication than other types of remote communication, such as that via telephone or virtual reality (Davis, Timpe-Laughlin, Gu and Ockey 2017, Qian 2009), the two modes are different and offer different experiences in communicating with others (Council of Europe 2018, Sellen 1995). Consequently, context validity of the VC mode cannot be warranted by administering tests in exactly the same way as in the face-to-face mode. Rather, adjustments have to be made to the administrative conditions and procedures of the VC tests (Clark and Hooshmand 1992), so as to ensure that the elicited language from candidates is of a comparable amount and quality to that of face-to-face tests.

Among the sparse literature that offers insights into the types of adjustments necessary for VC test administration, Clark and Hooshmand (1992) presented some useful suggestions based on post-test feedback from 30 examiners. In their comparative study of the face-to-face and VC modes of Arabic and Russian speaking tests administered at the Defense Language

Institute Foreign Language Center in the US, Clark and Hooshmand collected examiners' evaluation of their own examining performance and experiences of any difficulties during test administration. The VC tests were conducted using satellite-based video tele-training technology (VTT), and some difficulties were reported. Specifically, examiners experienced the sound dropping when both parties spoke simultaneously, the screen freezing, and distraction due to VC test procedures (i.e., changing tapes of a tape recorder). Consequently, Clark and Hooshmand suggested improvements in examiner training, particularly in minimising simultaneous speaking so as to avoid audio drop-outs.

Although the study by Clark and Hooshmand (1992) was conducted over three decades ago and recent advances in technology have overcome the problem of audio drop-outs with VTT, communication via VC remains prone to technical issues. Such issues include delayed synchronisation of audio and video, exaggerated movement (depending on the distance between the person and the webcam), restricted view of gestures due to screen size and the self-image window, and less obvious cues for turn-taking (e.g., Kim and Craig 2012, Kern 2014, Wang 2006). It is clear that examiner training for the VC mode needs to consider and incorporate measures to handle these issues. In addition to the ways in which technical glitches should be dealt with, additional VC guidelines are necessary, for example, when examiners try to interrupt candidates or elicit more speech, since the cues for turn-taking are less communicable in the VC mode.

The context of the current study

IELTS Speaking test

The specific context of the research reported in this chapter is the VC version of the IELTS Speaking test (a high-stakes test used to judge readiness for functioning in higher education, training programmes or general social contexts). The test is conducted in an examiner–candidate interview format, lasting 11–14 minutes in total. It consists of three parts: Part 1 (Introduction and interview; 4–5 minutes), Part 2 (Individual long turn; 3–4 minutes, including 1 minute of preparation time), and Part 3 (Two-way discussion; 4–5 minutes). The examiner plays the dual role of interlocutor and rater. Candidates are assessed on their entire performance on the test on four criteria (Fluency and Coherence, Lexical Resource, Grammatical Range and Accuracy, and Pronunciation), each of which has a nine-band scale.

Video-conferencing research of the IELTS Speaking test

Since 2014, a three-phase mixed methods research project has been conducted in different parts of the world (UK, China, Argentina, Colombia, Mexico and Venezuela), in order to explore the construct comparability of the face-to-face and VC-delivered IELTS Speaking test and to examine the feasibility of the VC-delivered test (see Nakatsuhara et al (2016, 2017b) and Berry, Nakatsuhara, Inoue and Galaczi (2018) for full reports from each phase).

The first phase consisted of a small exploratory study in the UK, involving 32 candidates and four examiners who administered both face-to-face and VC modes of the test. The comparability of test scores between the two delivery modes was established through many-facet Rasch modelling (MFRM). The range of language functions elicited by the two modes was also shown to be comparable, although a few functions were more prominent in the VC mode, flagging the need to address some VC-specific phenomena (e.g., greater need for meaning negotiation). Examiners' and candidates' perceptions of both delivery modes were also examined, and the VC test was found to be positively received by both examiners and candidates.

Based on the findings of the first phase, training materials on conducting and taking the VC modes were developed for examiners and candidates in Phase 2 of the research. This study, which took place in China, was of a larger scale, with 99 candidates and 10 examiners, involving the same two modes as the first phase. The findings confirmed those from the first phase, and a bespoke platform for delivering the VC mode of the IELTS Speaking test was developed. The third phase of the project was conducted with 89 candidates and eight examiners in four countries in Latin America. Phase 3 involved only the VC test, investigating further validity aspects of the VC-delivered test and the usability of the specific VC platform for the IELTS Speaking test.

As described in the Introduction, this chapter attempts to fill the research gap in the literature on the contextual parameter of test administration by reporting the detailed contents of examiner training for VC delivery of the IELTS Speaking test. In particular, we focus on an evaluation of the effectiveness of the training using feedback from examiners from Phases 2 and 3, as well as examiners' test administration behaviours observed in Phase 2. In so doing, areas for further improvement are identified, which contribute to more general discussion of the administration aspects of the context validity of VC speaking tests.

Research questions

The research questions (RQs) underpinning this chapter are:

RQ1. What were examiners' perceptions of the effectiveness of the VC test training?

RQ2. In the VC tests, did examiners use facilitation techniques from the training?

RQ3. What aspects of examiner training and administration for the VC test require further development?

Methodology

Participants

A total of 18 certified IELTS Speaking Examiners participated in this study and underwent VC examiner training (see the section 'Examiner training' for a detailed description). Their experience as IELTS Speaking Examiners ranged from two to 22 years (Mean = 8.91, SD = 5.44). Of the 18 examiners, 10 underwent training in Phase 2, where they administered both face-to-face and VC modes of the IELTS Speaking test to a total of 99 university students in Shanghai, China. All examiners and candidates for both delivery modes were located in the same examination venue. The overall IELTS Speaking scores of the 99 students ranged from Band 1.5 to Band 8.5 (Mean = 5.60, SD = 1.07). In Phase 3, eight examiners undertook training, after which they conducted VC tests with 89 candidates. Each candidate took the VC test administered by an examiner located in another country. Four of the eight examiners conducted VC tests from Bogotá, Colombia with 20 candidates from Medellín, Colombia and 25 from Mexico City, Mexico. The other four carried out VC tests from Buenos Aires, Argentina with 44 candidates from Caracas, Venezuela. The candidates' overall IELTS Speaking scores ranged from Band 4.0 to 8.5 (Mean = 6.15, SD = 0.92).

Eight trained invigilators, who also served as observers, took part in Phase 2. They were all PhD Applied Linguistics students from a university in Shanghai, China. Part 3 involved four trained invigilators, who were British Council personnel based in Mexico. Their roles and the training they received are detailed in the sections 'VC test set-ups' and 'Data collection and analysis'.

VC test set-ups

In Phase 2, the VC mode was administered using Zoom (Zoom Video Communications 2016), which offers high-definition VC and desktop sharing. Zoom was chosen because it was considered a more stable computer-mediated communication software than other programs.

An invigilator was present in all the VC candidate rooms in order to check candidates' IDs and to hand out and retrieve the topic card, pen and paper required in Part 2 of the test for note-taking, which examiners did in face-to-face test sessions. For the purpose of the research, there was an observer in each VC examiner room and face-to-face room. As explained in the section

'Data collection and analysis', the observers took notes of salient examiner behaviours to be analysed for investigating RQ1.

In Phase 3, the platform used was a Virtual Meeting Room (VMR) developed by Polycom and supplied by Videocall. Each test was essentially a virtual meeting with the examiner acting as the host and the candidate as the recipient of a meeting invitation. The crucial feature of this platform was its file-sharing facility, which allowed the examiner to display the topic card for Part 2 of the test on screen for the candidate, so as to eliminate the risk of invigilators handing a wrong topic card and to minimise potential time loss. Examiners chose one topic from the pool of (previously uploaded) topic cards and shared it with the candidate, removing it when Part 2 was finished. However, an invigilator was present in all VC candidate rooms in order to hand out a pen and paper during Part 2 of the test and to report any issues to the stand-by IT personnel if necessary. Additionally, during Part 2 of the test, special care was taken to keep a small window showing the examiner on the candidate screen so that candidates were fully aware of their listener/audience while producing monologic speech. Figure 1 shows the VMR test interface.

Examiner training

In both Phases 2 and 3, a one-day examiner training session for administering and rating VC-delivered tests was conducted by an experienced examiner trainer, who developed the training materials collaboratively with the researchers based on the Phase 1 study. The main points covered in the examiner training are presented in Figure 2 and are described in more detail in the following sections.

Step 1: Overview of the VC version of the IELTS Speaking test

On the training day, the trainer briefed the examiners about the purpose of the research, and informed the examiners of:

- what does not change in the VC mode
- what changes in the VC mode
- the room set-up for both face-to-face and VC tests
- the role of the invigilator/observer.

First, examiners were told that the examiner scripts remained the same in the VC mode, except for the addition of a preamble before the test and some minor changes in Part 2 instructions to indicate that the topic card, pen and paper are handed out and retrieved by an invigilator. The preamble consisted of a greeting and some simple questions (such as 'What time did you leave home today?'), which enabled a sound-check and gave both parties the opportunity to familiarise themselves with each other's voices. At the end of the preamble,

Figure 1 The bespoke VMR platform

Figure 1a Examiner screen for selecting a Part 2 topic card

Figure 1b Candidate screen displaying a Part 2 topic card and examiner window

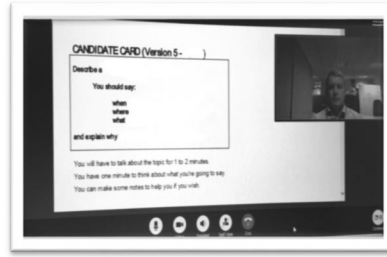
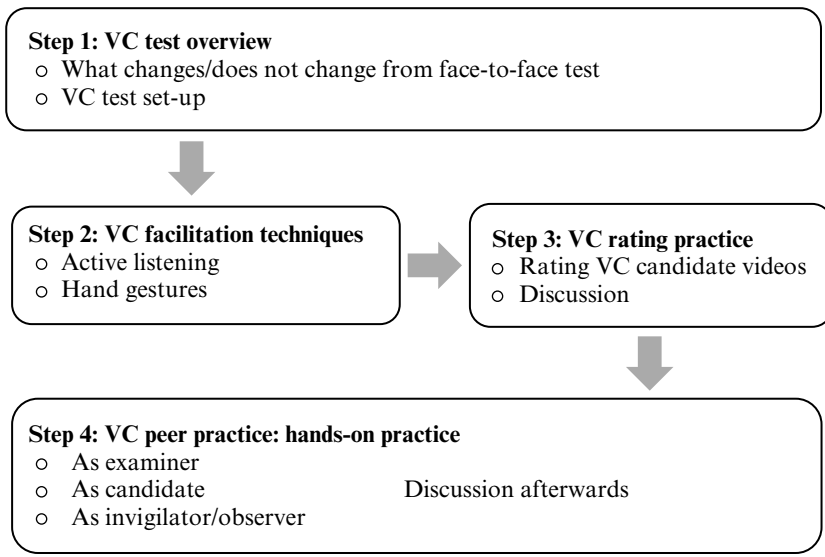


Figure 2 Overview of VC examiner training



the examiner was asked to remind the candidate to speak clearly at all times into the microphone, and then announce the start of the test.

For Phase 2, the room set-up for both modes was explained indicating that there would be an invigilator/observer in every room (i.e., face-to-face room, VC examiner room and VC candidate room), while Phase 3 only had an invigilator in each VC candidate room.

Step 2: Facilitation techniques for VC tests

Echoing the findings in Kern (2014), the trial of the VC mode in Phase 1 (Nakatsuhara et al 2016, 2017a) revealed that there were many cases of

delayed transmission of either the sound or video image, which made it difficult for examiners to back-channel or interrupt candidates where necessary. Therefore, in the Phase 2 study, briefing examiners about using non-verbal techniques instead of, or in addition to, verbal cues in order to facilitate the smooth running of tests was a crucial part of the training. Specifically, examiners were advised to:

- demonstrate active listening throughout all parts of the test by smiling and nodding to encourage candidates
- use hand gestures to stop or interrupt candidates if necessary, in addition to giving verbal instructions
- use hand gestures when trying to elicit more monologic speech in Part 2 from candidates who finished early, inviting further speech by gesturing to 'keep on going'.

Step 3: VC rating practice session

After receiving training, examiners participated in practice sessions to rate candidate performances. They were advised to apply the rating descriptors to performances as they usually do in face-to-face tests, and they then rated video-recorded performances of candidates from Phase 1 (in Phase 2 training) and Phase 2 (in Phase 3 training) and discussed the ratings together. The former included recordings of candidates at Bands 5.5, 6.0 and 6.5, the latter involved those at Bands 4.0, 6.5 and 7.0.

Step 4: VC peer practice sessions

Examiners then formed groups of three or four and took turns to practise administering (as examiner), test-taking (as candidate) and observing (as invigilator/observer) the VC tests. This also served as a hands-on practice session for the use of the test delivery platform of the respective phase. The practice tests were video-recorded for subsequent review. Examiners noted questions and difficulties that arose during these role-play sessions, and one practice test from each group was watched and discussed afterwards. After going through all four main stages of VC training, examiners completed an examiner post-training feedback questionnaire.

Data collection and analysis

The study involved four instruments: (a) examiner post-training feedback questionnaire, (b) examiner post-test feedback questionnaire, (c) observation notes, and (d) examiners' focus group discussions. The third instrument (observation notes) was applicable only to Phase 2 of the study, where the behaviours of VC examiners were compared with those of face-to-face examiners, but the rest of the instruments were common in both phases.

Examiner post-training feedback questionnaire

A 5-point Likert scale questionnaire was given to all examiners at the end of the training day, consisting of 10 questions about the levels of perceived usefulness of the training. Responses were collected immediately after the training had finished. Descriptive statistics were calculated on closed questions, and examiner comments in the free comment box were used to gather more detailed feedback for improving the training.

Examiner post-test feedback questionnaire

A 5-point Likert scale questionnaire was given to each of the examiners after they finished examining all candidates that were allocated to them. The questionnaire asked about the usefulness of training, as well as levels of comfort, and confidence in or ease of administering and rating the VC tests. Descriptive statistics were calculated on closed questions, and examiner comments in the free comment boxes were used to collect more detailed feedback for improving the training.

Observers' notes (Phase 2 only)

While the above questionnaires served to offer examiners' self-reported views of the effectiveness of the training and their own behaviours, self-reported data cannot be free from a potential misrepresentation due to the bias in the person providing the data (Baldwin 2000). In light of the exploratory nature of this study, we decided to obtain additional observational notes from a third party as triangulation data to uncover unconscious examiner behaviour that might be noticeable to the candidate.

The eight PhD Applied Linguistics students recruited from a university in Shanghai underwent training with one of the researchers on behaviour observation and note-taking and were briefed about the types of examiner behaviour to look out for. They were also asked to record any disruptive issues (such as loud background noise and technical failure) during a test session. To ensure uniformity of the noting format, a template was provided for taking notes (see Figure 3). Observers were

Figure 3 Example observation notes

Day	Obs ID	Cand ID	Mode	Test Version	Part 1	Part 2	Part 3	General Comments	Audio quality
4	20	S79	VC	3	Lots of gestures to help the C understand the Q; Very encouraging big smiles and nods whenever C talks.	Sat back as he asked the C to start speaking.	Spoke at a very slow rate intentionally; kept paraphrasing Qs using basic language.	Very smooth communication, no technical issues at all.	Very clear; No obvious impact on candidates' performance

allowed to use both English and their L1 (Chinese) as they wished, but were asked to type up their handwritten notes in English and send them to the researcher on the same day as the test. Example test notes are presented in Figure 3.

For each examiner–candidate pair, the same observer took notes in the face-to-face room and in the VC examiner room in order to ensure consistency of the observations of examiner behaviour. The eight observers/invigilators produced a total of 297 observation notes (99 candidates x three observations from the face-to-face, VC examiner and VC candidate rooms) with additional general comments.

All notes were collated and put into an Excel datasheet. NVivo Version 11 (QSR International 2016) was then used to organise the coding of what types of examiner behaviour were observed in both delivery modes. One of the researchers used emergent thematic coding to explore the categories of examiner behaviour, while attending to the VC facilitation features that examiners had been taught (see the section 'Step 2: Facilitation techniques for VC tests').

For coding the focus group discussions and observers' notes, approximately 10% of the data was firstly co-coded by two researchers, who reached an agreement rate of over 92%. After the coding reliability was established through partial double-coding and all discrepancies were fully discussed, one researcher carried out the rest of the coding.

Examiner post-test focus group discussions

After finishing all their test sessions and completing the examiner post-test feedback questionnaire, examiners discussed their experiences, views and issues regarding the VC tests in focus groups. As the data collection took place over several days in both phases, there were three focus groups in Phase 2, and four in Phase 3. The discussions were semi-structured and designed to achieve further elaboration of the responses in the examiner feedback questionnaires. All focus group discussions were recorded and transcribed for thematically coding the potential areas for improvement in the examiner training and test administration, except for one from Phase 3 that could not be retrieved due to a technical failure.

To summarise, Table 1 presents the relevant parts of the methodology in Phases 2 and 3.

Results and discussion

The results for each RQ (see the section 'Research questions') are discussed in each of the following sections.

Table 1 Overview of methodology in Phases 2 and 3

		Phase 2	Phase 3
Location(s)		Both examiners and candidates in Shanghai	Examiners in Bogotá and Buenos Aires; candidates in Medellín, Mexico City and Caracas
No. of participants	Examiners	10	8
	Candidates	99	89
	Observers	8	N/A
	Invigilators		4
Test mode(s) administered	Face-to-face	✓	N/A
	VC	✓	✓
VC mode arrangements	Technology used for VC tests	Zoom	Virtual Meeting Room
	Preamble given prior to test	✓	✓
	Part 2 topic card was shown to candidates by:	Invigilator handed it out and retrieved it	Examiner displayed it on screen and later removed it
	Paper and pen for Part 2 preparation time were given to and retrieved from candidates by:	Invigilator	Invigilator
Data sources for RQs	Examiner post-training feedback questionnaire (RQ1, 3)	✓	✓
	Examiner post-test feedback questionnaire (RQ1, 3)	✓	✓
	Observers' notes (RQ2)	✓	N/A
	Examiner post-test focus group discussions (RQ3)	✓	✓

Perceived effectiveness of VC examiner training

Table 2 summarises the 18 examiners' feedback obtained on the training day in Phases 2 and 3. As can be seen from Table 2, responses about the training day were very positive, with mean values between 4 (agree) and 5 (strongly agree) for all statements. It was particularly notable that all examiners strongly agreed to the usefulness of the training, the clarity of the explanation on the differences between face-to-face and the VC tests, and the availability of thorough discussion of VC facilitation techniques, indicating the successful provision of the training programme.

Table 2 Perceived effectiveness of VC training after the training day (N = 18)*

	Statement	Mean	SD
Q1	I found the training session useful.	5.00	0.00
Q2	The differences between the standard face-to-face test and the VC test were clearly explained.	5.00	0.00
Q3	What the VC room will look like was clearly explained.	4.46	0.63
Q4	VC facilitation techniques (e.g., use of preamble, back-channelling, gestures, how to interrupt) were thoroughly discussed.	5.00	0.00
Q5	The rating procedures in the VC test were thoroughly discussed.	4.83	0.38
Q6	The training videos that we watched together were helpful.	4.83	0.38
Q7	The peer practice sessions were useful.	4.83	0.51

**Note: 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree.*

Table 3 summarises the descriptive statistics from the statements from the examiner post-test feedback questionnaire that are relevant to examiner training. In this questionnaire, examiners indicated the extent to which the training had been useful during test delivery and rating.

As can be seen from Table 3, all statements presented mean values over 4 (agree), which indicates a high degree of agreement on the effectiveness of the VC examiner training. However, some statements received slightly varied responses according to the SDs of around 1, such as those regarding the handling of examiner scripts (Q4) and rating (Q5, Q6, Q7). Examiners' views and experiences in these aspects of the VC mode were further explored in the focus group discussions to identify areas for improvement in the training and administration of the VC tests (see the section 'Areas for improvement in VC examiner training and test administration').

Observed use of facilitation techniques for VC tests

As mentioned in the section 'Examiner training', examiners were specifically trained to use non-verbal techniques instead of, or in addition to, verbal cues in order to facilitate the smooth delivery of the VC tests.

Table 4 presents an overview of the 10 relevant coding categories of examiner behaviour in both delivery modes, in descending order according to the total frequencies of observed behaviour in the VC tests. For each part of the test, each cell represents the total number of test sessions across the 10 examiners where each type of examiner behaviour was noticed as salient behaviour and recorded by the observers. The total (the last column)

Table 3 Perceived effectiveness of VC training after live tests (N = 18)*

	Statement	Mean	SD
Q1	Overall, the examiner training adequately prepared me for administering the VC test.	4.67	0.59
Q2	The examiner training adequately prepared me for administering …….. of the VC test.		
Q2.1	• Part 1	4.78	0.54
Q2.2	• Part 2	4.61	0.49
Q2.3	• Part 3	4.78	0.54
Q3	I found it easy to handle topic cards on the screen in Part 2 of the VC test. [Phase 3 only]	4.63	0.74
Q4	The examiner training gave me confidence in handling the examiner scripts in the VC test.	4.72	0.95
Q5	Overall the examiner training adequately prepared me for rating test-taker performance in the VC test.	4.50	0.86
Q6	The examiner training adequately prepared me for applying the …… scale in the VC test.		
Q6.1	• Fluency and Coherence	4.56	0.78
Q6.2	• Lexical Resource	4.56	0.78
Q6.3	• Grammatical Range and Accuracy	4.56	0.78
Q6.4	• Pronunciation	4.39	0.78
Q7	The examiner training gave me confidence in the accuracy of my ratings on the VC test.	4.33	0.97

**Note: 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree.*

indicates the total number of test parts where the relevant behaviour was noted. Values for the face-to-face (F2F) mode are also listed, serving as benchmarks with which VC numbers are compared.

The most frequently noted examiner behaviour was found to be nodding, followed by the use of gestures and body language. Use of other paralinguistic features such as smiling and good eye contact (i.e., looking at the candidate on the screen) was also frequently noted. The ways in which the examiners responded to candidates' clarification requests were also found to be marked by the observers.

As can be seen in the last column of Table 4, the total number of test parts where the relevant examiner behaviour was observed as salient was consistently higher in the VC mode than in face-to-face. The same trend was found by test part, except for a few types of behaviour in certain test parts (e.g., uses response tokens in Parts 1 and 2). This was expected because, as noted in the section 'Step 2: Facilitation techniques for VC tests', the examiner training included facilitating techniques for managing the tests in the VC mode, based

Table 4 Overview of observed examiner behaviour (Phase 2)*

Types of examiner behaviour	Part 1		Part 2		Part 3		Total	
	F2F	VC	F2F	VC	F2F	VC	F2F	VC
Nods	32	41	31	39	19	27	82	107
Uses gestures and body language	19	24	13	21	31	40	63	85
Smiles	25	27	14	20	18	21	57	68
Responds to clarification request	14	20	3	0	15	20	32	40
Makes good eye contact	16	17	6	16	9	6	31	39
Tries to elicit more speech	0	1	10	12	6	7	16	20
Steers off-topic response	1	3	0	2	4	12	5	17
Speaks differently compared to the other mode	3	6	1	2	2	6	6	14
Uses response tokens	1	0	1	0	4	10	6	10
Stops/Interrupts candidate	1	9	0	2	2	1	3	12

**Notes: F2F = face-to-face. The maximum number of sessions for each cell is 99 for each test part and 297 for total.*

on insights from the literature (Clark and Hooshmand 1992, Kern 2014, Wang 2006) and the Phase 1 study of the project (Nakatsuhara et al 2016, 2017a). Examples of observer notes on the use of these techniques are shown below. Please note all observer notes are unedited to maintain authenticity.

- 'When the examiner was delivering the instruction, he used a lot of hand gestures. For example, he lifted one finger when he said "one minute". […] Examiner nodded all the time while hearing the candidate's answering.' (Examiner E, Candidate S025, Part 2)
- 'The examiner used hand gesture once when explaining the question to the candidate. […] When hearing the answer, he nodded and said "uh huh" to echo with the candidate.' (Examiner B, Candidate S007, Part 3)

Some examiners were also observed to speak slightly differently between the face-to-face and VC modes (i.e., under the category 'Speaks differently compared to the other mode' in Table 4), e.g., speaks in a louder voice and emphasises some key words more in the VC mode. Furthermore, some observers' notes were coded under categories of 'Tries to elicit more speech' and 'Stops/interrupts candidate'. Both types of examiner behaviour correspond with the strategies featured in the examiner training for extending the long turn in Part 2 (if necessary) and interrupting candidates effectively in the VC mode (see the section 'Step 2: Facilitation techniques for VC tests'). The use of techniques to 'steer off-topic responses' (such as raising a hand and doing a 'stop' sign) were also used more often in the VC than in the face-to-face mode. In short, in response to RQ2 it appears clear from observers' notes that the examiners were, in general, using the taught techniques and strategies effectively during the VC tests.

It is important to note that a lack of observations in the observers' notes does not necessarily mean the absence of a behaviour. As noted earlier, the observers were instructed to note the salient features of examiner behaviour, and there may still have been different degrees of saliency or noticing among the observers. Nevertheless, the provision of training and the arrangement for the same observer to observe an examiner across both test modes are believed to have warranted a satisfactory level of objectivity in the observers' notes.

Areas for improvement in VC examiner training and test administration

Transcripts of the focus group discussions from Phases 2 and 3 were thematically coded. Four relevant coding categories and excerpts are reported in this section in order to answer RQ3, which concerned the potential areas for improvement in the VC examiner training and test administration. For the full lists of coded categories, see Nakatsuhara et al (2017b) for Phase 2 and Berry et al (2018) for Phase 3.

Training on VC rating

In the training for both Phases 2 and 3, there were three practice sessions on rating (using the video recordings from previous phases) and three practice sessions on administering the test (with a fellow examiner) under the VC mode. Commenting on the range of test-taker ability assessed in the training, one examiner noted:

> As you've seen just from these five days [of data collection] where one person had a candidate that was Band 1, I had one that was a Band 2.5… So what we did in the training was that we saw a video of candidates that were round about 4.5 to 7, which is a good range, and we practised with each other. But it would be good if the trainees got to see a video of a very low candidate possibly also a very high one as well. (Examiner J, Phase 2)

Some examiners agreed with Examiner J's comment and suggested the inclusion of a wider range of ability levels with a very low-level candidate (i.e., Band 2.5 or lower) and a higher-ability candidate (i.e., Band 7.5 or higher). Other examiners commented that it might be beneficial for them to have practice sessions with prospective candidates, rather than peer examiners, where they practise rating while conducting the interview in the VC mode, so as to make them more confident in their ability to rate when the VC training finished.

VC-specific guidelines for Part 2

Several comments were made in the focus group discussions in relation to the need for more specific examiner guidelines for Part 2 in VC test administration. The first issue was that some of the examiners were unsure about what to do while candidates were preparing for Part 2, which includes one minute of preparation time: 'I think there have got to be more guidelines about how to deal with Part 2 preparation because just staring at the candidate will freak things out' (Examiner A, Phase 2). Other examiners made suggestions as to what they could do during Part 2 preparation: 'It's a bit like newsreaders. They are told to shuffle papers around. So I think we need that' (Examiner C, Phase 2). By employing a 'newsreader technique', examiners would look down and shuffle papers around in Part 2, in the same way that newsreaders do when they are on camera but not reading the news. This might reduce awkwardness while examiners wait for the candidates to finish their preparation.

Another issue reported by some examiners related to the occasional difficulty of eliciting further speech in Part 2 under the VC mode:

> I had a candidate today who stopped early in Part 2 […] and I had no way of getting her to continue except the rounding-off questions. And

> because she stopped so early we got through the rounding-off questions and we still hadn't quite reached two minutes. When I did her in the face-to-face she did the same thing, I was able to get her to go more by pointing at one of the questions on the cue card. (Examiner J, Phase 2)

> In Part 2, when a candidate gets stuck and has not covered some of the prompts, the examiner cannot point at the prompts not used [as in face-to-face], for example. The only possible help is [to ask] "Can you tell me more about…?" Would it be possible to include some back-up prompts in the script? (Examiner K, Phase 3)

Despite engaging in active listening and using hand gestures to elicit further speech from candidates, as per VC examiner training, examiners from both phases raised the concern that in the VC delivery mode it is still quite challenging to facilitate candidates to speak more when they finish early in Part 2. In face-to-face tests, examiners can just silently point to the bullet points on the topic card without deviating from the examiner scripts. However, this is not possible in the VC tests, and there is clearly a limit to the range of non-verbal cues available for examiners in this mode. In the VC tests, therefore, examiners may benefit from having the option of using additional verbal cues to elicit more speech, should they become necessary. Such optional additions to the examiner scripts should be used in a principled manner, so that they do not pose a threat to the scoring validity of the test – examiners would be able to elicit sufficient amounts of language in Part 2, but at the same time, not give extra scaffolding and an advantage to some candidates.

Extension of VC Part 2 time

The third area for potential improvement, also related to Part 2, concerned the timing of Part 2 of the test. All examiners in Phases 2 and 3 agreed that the four minutes allocated for Part 2 was too short in the VC tests. Comments from examiners relating to this aspect of the VC delivery included:

> I'm not asking rounding-off questions, it's the only way to keep it within the four minutes […] if we're going to deliver this as scripted and there's no modifications if we want to deliver the rounding-off questions on a regular basis then I think certainly for this there needs to be an extension of time for Part 2 of about thirty seconds and make it about four and a half minutes for this version. (Examiner N, Phase 3)

In the VC tests in Phase 2, examiners had to wait while the invigilator handed and retrieved the topic card, paper and pen to/from the candidate (Nakatsuhara et al 2017b). Similarly, in Phase 3, where a bespoke VC platform with file-sharing facility was used, extra time was necessary to display the topic card on the screen, wait for the invigilator to hand the paper

and pen to the candidate for note-taking, and wait for the candidate to give back the paper and pen to the invigilator after the monologue (Berry et al 2018). The consequence of running late in Part 2 would be a reduced amount of time available for the rest of the VC test. Accordingly, a suggestion was made by some examiners to extend Part 2 of the VC tests by 30 seconds.

Troubleshooting guidelines

The fourth area that examiners suggested needed improvement related to when something goes wrong during the VC test. Depending on the timing and duration of the trouble, all examiners agreed that it may be useful to have guidelines which include information regarding when to continue or go back and re-do part(s) of the test.

> Just thinking about the what ifs, … on Saturday [the screen] froze for half a minute, … it was at four minutes thirty – the last question, just at the start of it. [So] it wasn't that overly important. But if that is going to happen, for example, halfway through Part 1, Part 2, what are the rules, what should one do; should one go back and start it all over again and hope for the best? (Examiner E, Phase 2)

> I know I could fix those problems we had today [by rebooting the platform] but when I was in that room examining I had one priority and that was that candidate. And when it failed I wanted [the IT support personnel] to take over the issue of technology because my priority was the candidate in front of me. If then [I] say "I'll deal with this", I'd feel I wouldn't be giving the appropriate attention to the candidate. (Examiner N, Phase 3)

Troubleshooting guidelines would also need to specify when and how to call for IT support outside the VC test room (and how to communicate it to the candidate). Although examiners could be given an option to reboot the VC system to fix the problems themselves, it would divert their attention from the assessment of the candidate. As Examiner N said, it seems critical to have IT support on stand-by for VC tests.

Conclusions

As mentioned in the Introduction, the aim of this study was to investigate the administrative setting parameter of the VC mode of the IELTS Speaking test, which falls under the remit of context validity (Taylor 2011, Weir 2005). Test administration procedures are an integral aspect of context validity, since they ensure that candidates are given appropriate and equitable opportunities to demonstrate the expected target language performance. While this was a relatively small-scale study, the findings indicated that the examiners found the VC training very useful (see the section 'Perceived effectiveness of VC

examiner training'), and the observers reported a number of cases where facilitation techniques and strategies taught in the VC examiner training were used effectively (see the section 'Observed use of facilitation techniques for VC tests'). Additionally, the focus group discussions revealed a number of issues of importance for the operationalisation of a VC IELTS Speaking test (see the section 'Areas for improvement in VC examiner training and test administration'), which have valuable implications for VC tests in general.

Ensuring the context validity of the VC mode of a currently face-to-face speaking test is not achieved by replicating exactly the same procedures and timings of the face-to-face test, but by ensuring an equally rateable sample of language is elicited as in the face-to-face mode, thus leading to comparable scores. The VC mode involves potentially obstructive features, just as real-life VC communication does, such as delayed synchronisation of image and sound and less obvious cues for turn-taking (Kern 2014). Thus, if minor changes are required to achieve the comparability of opportunities for candidates to demonstrate their ability, they should be regarded as necessary and justified modifications to enhance the comparability of the two test modes, rather than as a threat to test validity. What is crucial is that such changes in the administrative conditions need to be considered at the initial stages of designing the VC mode and are supported by evidence.

One area, which was not focused on in this chapter but needs careful consideration when introducing the VC mode of a face-to-face test, is practical arrangements for note-taking. If a speaking test requires the candidate to take notes, due consideration must also be given to the room set-up, especially regarding whether to have an invigilator in the VC candidate room and, if so, deciding what role(s) they will play during the test. Having an invigilator may be important from a security and malpractice point of view, and in case technical problems occur. If invigilators are to hand out and retrieve paper and pen/pencil used during the preparation time of a monologic task, the timings need to be further standardised. If, on the other hand, invigilators are not part of the test administration, alternative arrangements must be in place to enable candidates (and examiners) to report any problems during the test session, and to manage paper and pen/pencil note-taking. It is worth considering that instead of using paper and pen/pencil, an alternative way of taking notes for the test tasks could be to harness PC/tablet capabilities for note-taking on screen. This would, of course, come with a number of practical issues to be resolved, such as screen size and provision of tablet pens, and validity concerns such as candidate familiarity with this form of note-taking.

Following this project, an operational trial was recently conducted as Phase 4 of the project (Lee, Patel, Lynch and Galaczi 2021). The study focused on timing in the VC mode, seeking empirical evidence for the suggested extension of timing for Part 2 of the test reported in the section 'Extension of VC Part 2 time'. Such an iterative cycle of evidence collection

in a four-phase design provided depth and breadth of evidence to support the VC mode of the IELTS Speaking test.

In summary, by explicating the administrative details of the VC-delivered IELTS Speaking test and by discussing the analysis of examiner feedback and observers' notes, this chapter has highlighted a number of practical areas which test providers should consider when a new VC delivery mode of a speaking test is introduced. As stated throughout, a change of test delivery modes cannot be validly achieved without careful consideration of various training and administrative features, which ultimately impacts on test validity. The findings reported here come from one test only but have broader implications for VC speaking tests in general in providing insights to inform their practical design and theoretical discussions about test validity. While further technological innovations are likely to offer a wider range of test delivery modes in the future, we should be reminded of the importance of having clear, appropriate procedures and guidelines for training and test administration in place to ensure the quality of a test.

Acknowledgement

This project was funded and supported by the IELTS Partners, and we would particularly like to thank Mina Patel, Val Harris and Sonja Webb for their invaluable contributions.

References

Baldwin, W (2000) Information no one else knows: The value of self-report, in Stone, A A, Turkkan, J S, Bachrach, C A, Jobe, J B, Kurtzman, H S and Cain, V S (Eds) *The Science of Self-report: Implications for Research and Practice*, Mahwah: Lawrence Erlbaum Associates, 3–7.

Bernstein, J, Van Moere, A and Cheng, J (2010) Validating automated speaking tests, *Language Testing* 27 (3), 355–377.

Berry, V, Nakatsuhara, F, Inoue, C and Galaczi, E D (2018) *Exploring the use of VC technology to deliver the IELTS Speaking Test: Phase 3 technical trial*, IELTS Partnership Research Papers 2018/1.

Brown, A (2003) Interviewer variation and the co-construction of speaking proficiency, *Language Testing* 20 (1), 1–25.

Clark, J L D and Hooshmand, D (1992) "Screen-to-screen" testing: An exploratory study of oral proficiency interviewing using video teleconferencing, *System* 20 (3), 293–304.

Council of Europe (2018) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with new descriptors*, Strasbourg: Council of Europe, available online: rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Davis, L, Timpe-Laughlin, V, Gu, L and Ockey, G J (2017) Face-to-face speaking assessment in the digital age: Interactive speaking tasks online, in Davis, J M, Norris, J M, Malone, M E, McKay, T H

and Son, Y A (Eds) *Useful Assessment and Evaluation in Language Education*, Washington, D.C.: Georgetown University Press, 115–130.
Galaczi, E D (2010) Face-to-face and computer-based assessment of speaking: Challenges and opportunities, in Araujo, L (Ed) *Computer-based Assessment of Foreign Language Speaking Skills*, Luxembourg: Publications Office of the European Union, 29–51.
Kern, R (2014) Technology as *Pharmakon*: The promise and perils of the internet for foreign language education, *The Modern Language Journal* 98 (1), 340–357.
Kim, J and Craig, D A (2012) Validation of a videoconferenced speaking test, *Computer Assisted Language Learning* 25 (3), 257–275.
Lazaraton, A (2002) *A Qualitative Approach to the Validation of Oral Language Tests*, Studies in Language Testing Volume 14, Cambridge: UCLES/ Cambridge University Press.
Lee, H, Patel, M, Lynch, J and Galaczi, E D (2021) *Development of the IELTS Video Call Speaking Test: Phase 4 operational research trial and overall summary of a four-phase test development cycle*, IELTS Partnership Research Papers 2021/1.
Nakatsuhara, F (2008) Inter-interviewer variation in oral interview tests, *ELT Journal* 62 (3), 266–275.
Nakatsuhara, F, Inoue, C, Berry, V and Galaczi, E D (2016) *Exploring performance across two delivery modes for the same L2 speaking test: Face-to-face and computer delivery using VC technology: A preliminary comparison of test-taker and examiner behaviour*, IELTS Partnership Research Papers 1.
Nakatsuhara, F, Inoue, C, Berry, V and Galaczi, E D (2017a) Exploring the use of VC technology in the assessment of spoken language: A mixed-methods study, *Language Assessment Quarterly* 14 (1), 1–18.
Nakatsuhara, F, Inoue, C, Berry, V and Galaczi, E D (2017b) *Exploring performance across two delivery modes for the IELTS Speaking Test: Face-to-face and VC delivery (Phase 2)*, IELTS Partnership Research Papers 3.
O'Loughlin, K (2001) *The Equivalence of Direct and Semi-direct Speaking Tests*, Studies in Language Testing Volume 13, Cambridge: UCLES/Cambridge University Press.
Qian, D (2009) Comparing direct and semi-direct modes for speaking assessment: Affective effects on test-takers, *Language Assessment Quarterly* 6 (2), 113–125.
QSR International (2016) *NVivo Version 11*, available online: qsrinternational.com
Sellen, A J (1995) Remote conversations: The effects of mediating talk with technology, *Human-Computer Interaction* 10 (4), 401–444.
Taylor, L (2011) Introduction, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing Volume 30, Cambridge: UCLES/Cambridge University Press, 1–35.
Wang, Y (2006) Negotiation of meaning in desktop videoconferencing-supported distance language learning, *ReCALL* 18 (1), 122–146.
Weir, C J (2005) *Language Testing and Validation: An Evidence-based Approach*, Basingstoke: Palgrave Macmillan.
Xi, X (2010) Automated scoring and feedback systems: Where are we and where are we heading?, *Language Testing* 27 (3), 291–300.
Zoom Video Communications (2016) *Zoom version 3.5.16903.0522* [Application software], available online: www.zoom.us

11 The interface between diagnostic writing assessment systems and a socio-cognitive validity framework

Stephanie Link
Oklahoma State University, USA

Abstract

Computer-based diagnostic writing assessment systems have contributed to language testing by providing rapid measurement of student knowledge and language development. However, as technology advances there is a need to keep pace with theory development and theory testing. This chapter argues that by integrating theory at the onset of design and validation of automated assessment systems, a stronger link can be made between how feedback is generated and the ways in which feedback is used and interpreted for diagnostic purposes in the language learning classroom. Furthermore, current understanding of the socio-cognitive processes involved in dynamic, nonlinear language development can be used to validate measures of the writing construct for producing simplified feedback models that account for the educational culture and social context of language use. This chapter suggests following a socio-cognitive approach to validation that can support future development of diagnostic systems using automated writing evaluation (AWE). I highlight critical questions and key validation components from Shaw and Weir (2007) and Weir (2005) and suggest *a priori* and *a posteriori* validity evidence that can be collected to assist stakeholders in making judgments about automated feedback use and interpretation for context-specific scenarios. I then introduce an evolving AWE diagnostic tool as proof-of-concept for how to collect scoring validity evidence that aligns theory and diagnostic practice.

Introduction

Advancements in technology have enabled researchers to employ methods from areas such as artificial intelligence and feature tracking in order to

implement more effective tools for assessing knowledge and enhancing the language performance of students. As methods become more complex and specialized, and algorithms expand the possibilities for gathering and analysing vast amounts of data, their potential for contributing to developments in diagnostic assessment grows. The heightened understanding of students' knowledge and development that efficient and effective computer-based diagnostics can provide lends itself to advancements in theories of language learning and instruction. Consequently, there is a more pronounced need for integrating theory to sufficiently advance methods of technology development in order to keep pace with theory development and theory testing and to maintain a coherence between theory and language learning diagnostics.

While the development of useful writing systems has long been a goal for computational linguists and software engineers serving our professional community of language testers and practitioners, this task entails building tools that are easy to use and that provide feedback that is easy to interpret. Otherwise, the use and interpretation of automated feedback risks being meaningful only to their developers, who will have a clear understanding of how algorithms were designed to represent a language construct. To become more widely accepted across assessment and classroom contexts, development and validation of writing assessment systems would benefit from closer ties to second language acquisition (SLA) and learning theories that tend to already inform teaching practices and learning activities.

In this chapter, I argue for future development and validation of diagnostic assessment systems that integrate automated writing evaluation and/or automated essay scoring[1] by employing a socio-cognitive lens to capture a humanistic view of writing as both a socially situated and a cognitively processed phenomenon (Shaw and Weir 2007, Weir 2005). With respect to the enduring challenges and opportunities that automated assessment has to offer, this paper provides a critical appraisal of the evidence needed to validate a new diagnostic assessment tool within a socio-cognitive framework for a target language situation and target test-takers. I then demonstrate how to collect and evaluate theoretical and empirical evidence at the onset of assessment tool development by introducing an evolving diagnostic system called CAFFite, that highlights critical attention towards data- and theory-driven diagnostic assessment of second language writing development as proof-of-concept.

1 In this paper, automated writing evaluation (AWE) refers to feedback tools that provide both diagnostic holistic scores and corrective feedback on writing samples. Automated essay scoring (AES) includes only a holistic score.

A socio-cognitive framework for developing and validating diagnostic systems

The firmly established era of computing and big data is beginning to mediate, augment, and regulate assessment processes as well as (re)define fundamental constructs of language performance. As diagnostic writing assessment systems using AWE and AES continue to evolve, there is a need to critically evaluate the empirical and theoretical contributions of automated feedback models and to push for advancements that deepen our understanding of language performance and development. There is thus a need for developers to be aware of the established theory relating to cognitive processing and real-life language use to help inform model development. Deane (2013:21) further argues that:

> The ultimate promise of automated writing evaluation emerges when we open up the space of possibilities and consider it not only as a technology that supports automated essay scoring, but as the basis for large-scale, embedded forms of automated writing analysis in which social and cognitive aspects of the writing process are taken more richly into account.

To align with Deane's call for future AWE research from a socio-cognitive approach, a discussion of critical questions surrounding an assessment's target use situation and test-taker characteristics is desirable. Such a dialogue can help to carefully position future technology development to meet the demands of validity and the operationalization of criterial distinctions used to construct the automated diagnostic feedback. Shaw and Weir (2007) outline key questions as part of their socio-cognitive validity framework that addresses the cognitive dimension of the writing process by testing the mental processing abilities of test-takers while giving equal weight to social uses of language (the social dimension). Their unified approach to validity reconfigures conventional sources of validity (construct, content, criterion) as an interaction between constituent parts, namely the traits of communicative language ability, the context of use, and the score (including its interpretation). More specifically, they argue for *a priori* (before-the-test event) and *a posteriori* (after-the-test event) validation components (i.e., context, cognitive, scoring, consequential, and criterion-related validity). These components can be translated to the context of before and after development of feedback algorithms for AWE and AES. Table 1 highlights the key questions aligning with each validation component and then illustrates how future models may seek to address these questions by collecting validity evidence in order to be confident that the nature and quality of automated feedback matches what is adequate and appropriate for target use and interpretation of feedback.

Table 1 Alignment of critical questions, validation components, and necessary evidence using a socio-cognitive validity framework for development and evaluation of computer-based diagnostic assessment systems using automated writing evaluation (AWE)

Critical questions[a]	Validation components	Necessary validity evidence about computer-based diagnostic assessment systems	
		a priori evidence	*a posteriori* evidence
Are the characteristics of the writing tasks and their administration appropriate and fair to the candidates who are completing them?	Context validity	– Domain analysis is conducted to identify linguistic, social, and cultural requirements of writing tasks within the target context – Survey of prospective users and their perceptions towards task administration and fairness without automated feedback	– Feedback engine is able to measure various genres identified in domain analysis with equivalent accuracy – Survey of users and their attitudes towards task administration and fairness of automated feedback – Statistical analysis of differences in user perceptions of writing tasks with and without automated feedback
Are the cognitive processes required to complete the writing tasks using automated feedback appropriate?	Cognitive validity	– Piloting and trialing task conditions and genres with optional retrospective interviews, think-aloud protocols, and observations – Collection of a representative sample of essays from each task condition and genre – Manual measurement of performance using validated measures of quality writing	– Automated measurement of performance using automated feedback – Statistical analysis of human vs computer scoring with varying task conditions and genres (optionally feature detection techniques such as eye-tracking or screen-capturing data)
How far can we depend on the feedback which results from the computer-based diagnostic system?	Scoring validity	– Feedback algorithms are trained to extract features from essays using validated measures of quality writing performance – Human–human reliability estimates are provided over occasions of assessment, versions of writing tasks, and with different performance levels	– Human–computer reliability estimates reflecting consistency that would appear across raters, over occasions of assessment and versions of writing tasks, with different performance levels, and across various learner groups – Precision/recall and linguistic analyses demonstrating accurate application of measures into feedback generation model

What external evidence is there outside of the automated feedback that the score is fair?	Criterion-related validity	– Statistical analysis of the relationship between automated feedback and external measures, such as the CEFR[b] language standards or ACTFL[c] performance descriptors
What effect does automated feedback have on various stakeholders?	Consequential validity	– Washback studies investigating social and cultural consequences of using the automated feedback and of the decisions made based on use for all stakeholders – Feedback reports are developed for stakeholders in an understandable way and in a timely fashion – Feedback is treated as confidential

[a] *Critical questions are adapted from Shaw and Weir (2007) to reflect computer-based diagnostic writing feedback use and interpretations.*
[b] *Common European Framework of Reference for Languages (Council of Europe 2001)*
[c] *American Council on the Teaching of Foreign Languages (ACTFL 2012)*

According to Shaw and Weir (2007), content validity is not only about linguistic content (e.g., lexical and structural resources, discourse mode) but also the social and cultural content in which a writing task is performed. That is, evidence should account for the setting under which a task is performed or an assessment is administered. For diagnostic assessment systems, validity evidence can be collected using domain analysis to determine the genres and task characteristics typically required in a given context. Perception-based research can uncover pre-existing concerns with the setting that diagnostic assessments can attempt to address. After development, a system can be tested for how accurately it can detect the linguistic features of various genres and what users' thoughts are towards the new setting.

Cognitive validity involves collecting evidence of the cognitive processing triggered prior to and while using a diagnostic system's automated feedback. Collection of this type of evidence requires that system developers and assessment administrators are aware of cognitive processing theory and how processing transfers to real-life language use. This understanding can help to select and implement appropriate methods (e.g., retrospective interviews, think-aloud protocols, and observations) for testing processing under different task conditions and while writing different genres both with and without automated feedback.

Scoring validity was chosen for the focus of this chapter because this validation component is an important part of construct validity since scoring describes the level of performance. In addition, much of the uncertainty about automated feedback can be traced back to lasting concerns about construct under- or misrepresentation (Deane 2013, Xi 2010). Scoring models also impact what diagnostic feedback the test-taker receives, thus forming a foundation for measuring the writing construct. Shaw and Weir (2007:143) define scoring validity as:

> … all aspects of the testing process that can impact on the reliability of test scores. It accounts for the extent to which test scores are based on appropriate criteria, exhibit consensual agreement in marking, are free as possible from measurement error, stable over time, consistent in terms of content sampling and engender confidence as reliable decision-making indicators.

More information of the validity evidence collected for demonstrative purposes will be discussed in the following section.

Shaw and Weir (2007) suggest that criterion-related validity is primarily an *a posteriori* concept. Evidence can be concurrent and predictive. This validity component concerns the extent to which feedback correlates with external measures of performance. External scales should be selected

based on the same theoretical foundation that grounds the development of feedback algorithms. Finally, consequential validity is used to determine whether the social impact of the computer-based assessment aligns with the assessment purpose and other social values, such as washback on classroom practices. Thus, evidence is also largely *a posteriori*.

Decisions made throughout the validation process will impact the type and amount of validity evidence needed to make effective judgments about validity. That is, the target use situation, the test-taker characteristics, and scoring criteria built into a system's feedback model will impact how the feedback is interpreted and used. While collecting validity evidence for each interrelated component of the framework is integral to evaluating the validity of an automated feedback engine for a specific use, the next section provides a proof-of-concept by focusing on scoring validity. To do this, I introduce the *a priori* and *a posteriori* evidence collected to develop and then validate a diagnostic assessment tool that provides data- and theory-driven understanding of second language writing development.

Data- and theory-driven development and validation: The case of scoring validity

While research on automated feedback algorithms has argued for the accuracy of the engines in relation to human scoring (e.g., Attali, Bridgeman and Trapani 2010, Burstein and Chodorow 2003, 2010), there is still skepticism about the validity of score use and interpretations. This uncertainty is partially due to limited justification for how and why language features are measured in a feedback model, which can lower stakeholders' trust in the automated feedback (Li, Link, Ma, Yang and Hegelheimer 2014). By documenting snapshots of how acquisition unfolds across time and task conditions, assessment tools can shift to a more dynamic score generation process that has the potential to not only provide scores for placement and proficiency but also for diagnostic assessment and classroom use. This section provides a proof-of-concept snapshot of the development and evaluation of an evolving AWE diagnostic assessment tool called CAFFite for measuring second language learners' writing development using measures of complexity, accuracy, fluency, and functionality (CAFF). The engine was conceptualized to provide holistic, summative feedback and precise measures of language development for more individualized, diagnostic accounts of writing profiles. First, the target use situation and test-taker characteristics are outlined. Then, specific focus is placed on how scoring validity evidence was collected so as to illustrate a blueprint for future validation research.

Target use situation and test-taker characteristics

The specific context of CAFFite use is a real-life academic context of an English-medium post-secondary institution in the USA. In this context, the English department offers non-native English-speaking undergraduate and graduate students advanced courses to support academic English skills. Two writing courses for undergraduates are concurrently offered (Level 1 and Level 2). Students begin at Level 1 in their first semester of enrollment based on results from an in-house English Placement Test and then proceed to Level 2 for their second semester; students may also be placed directly into Level 2, or they may be placed directly into first-year composition courses with native English speaking students (Level 3).

Level 1 is for students at intermediate-high proficiency (ACTFL 2017) who lack grammatical control and the ability to effectively convey meaning. Instruction in this class concentrates on the essentials of academic writing at sentence and paragraph level. Students practice grammar, vocabulary, mechanics, style, and organizational patterns, as well as the key compositional processes of planning, drafting, and revising. Level 2 is for advanced-low students who have a control of grammar but need additional support in cohesion, coherence, and organizational strategies. Instruction at this level focuses on writing professional communication, academic papers and reports, and in using published source material in writing. It provides experience in presenting oral reports and participating in discussions and prepares students for academic writing in their disciplines. After completion of Level 2, students are expected to be ready to enter into first-year composition (Level 3); that is, they are prepared for attending mainstream classrooms and writing in various academic genres.

The learners in this target domain can come from a wide range of multilingual language backgrounds and may include: (1) international students who arrive in the US in pursuit of an education, (2) resident immigrants who arrive as adolescents or young adults, and (3) children of resident immigrants who arrive at a young age or are born in the US (see Ferris (2011) for a description of learner groups). CAFFite was designed to assess these students' developmental trajectories when they pass into and through the various levels. Holistic scores as well as analytic scores provide them with individualized, diagnostic feedback to highlight specific strengths and weaknesses in their writing performance and provide teachers and administrators with descriptors of each student's needs.

Design and training of algorithms using valid measures of performance: *A priori* evidence

When considering how to collect evidence for demonstrating scoring validity, it is important to first collect *a priori* evidence that the diagnostic feedback algorithms are developed and trained to extract linguistics features using validated measures of quality writing performance. Attali and Burstein (2006), for example, developed a version of a feedback engine from Educational Testing Service (ETS) called e-rater. The developers acknowledged that automated feedback 'was always based on a large number of features that were not individually described or linked to intuitive dimensions of writing quality' (2006:7). They suggested a scaled-down approach that includes a small set of features that are the most important characteristics of quality writing, claiming that evaluating a small set of features allows for greater understanding and control over the automated scores. SLA theory merits added attention when developing and evaluating automated systems for diagnostic assessment and computer-assisted language learning (Chapelle 2001, 2009) because SLA can provide specialized insight into socio-cognitive processing that generic feedback models only begin to describe because of their basis in computation and programming.

To collect *a priori* evidence for scoring validity, CAFFite algorithms were created using validated measures of quality writing performance. These algorithms were first grounded in two theoretical perspectives – Complexity Theory (Larsen-Freeman 2006, Larsen-Freeman and Cameron 2008) and Systemic Functional Linguistics (SFL) (Halliday (Ed) 1979, Halliday and Hasan 1989). Both theories take a systems-based perspective towards language acquisition, and I argue that they are complementary in their notions of how language emerges and the relation between its emergence and the multifaceted learning environment (see Link (2015) for an expanded discussion). The first, Complexity Theory, posits that language development is a dynamic, non-linear process within a system. Although development is nearly impossible to predict, the types of behavior likely to occur can be mapped to 'possible and probable patterns' within a learner's specific system and all its interacting parts (Larsen-Freeman and Cameron 2008:17). The notion of 'complex' within a complex system means that the elements in a system are ever-changing and continuously adapting in response to feedback. SFL can support the understanding of complex language development by use of its functional, flexible grammar to analyze language features related to how meaning is constructed, used, and developed (Halliday and Matthiessen 2014), offering a principled basis for describing language variation in relation to both who is using the language and the purposes for which it is used (Halliday and Hasan 1989).

From these two theoretical perspectives, patterns of development can be visualized by use of carefully selected measures that can document changes in a

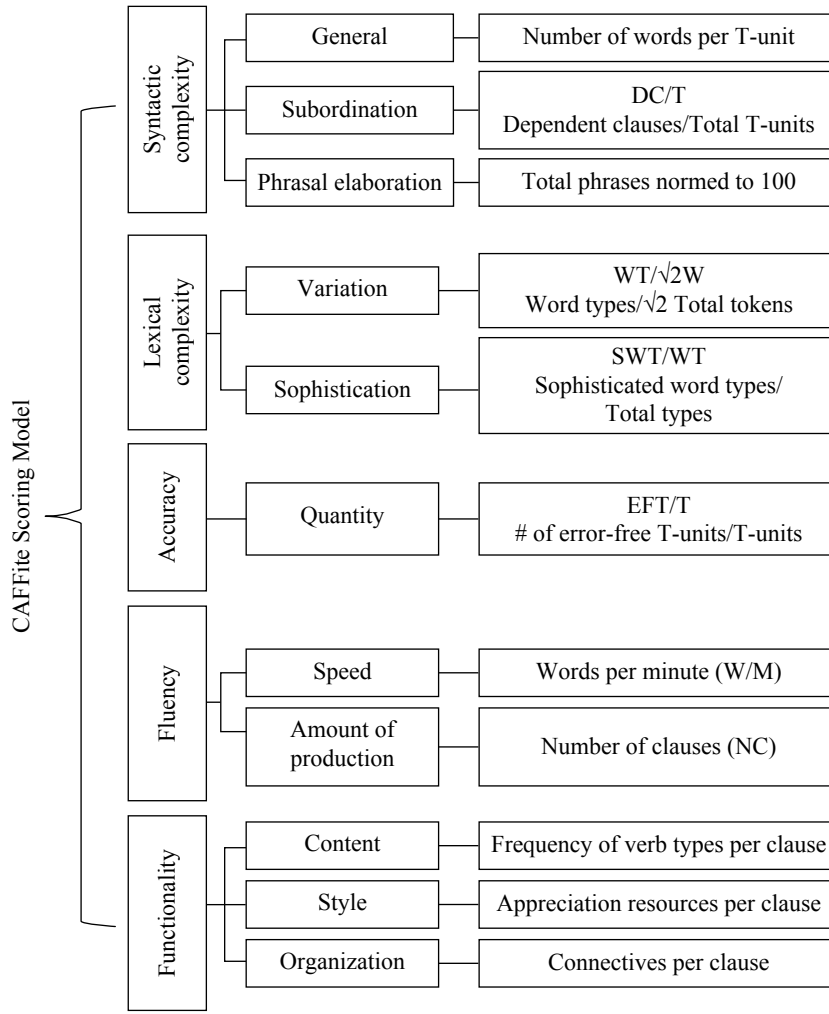
system as acquisition unfolds in order to better understand why change occurs or why not much change seems to take place. Selection of these measures also draws on the notion of construct multidimensionality in which multiple dimensions of proficiency are measured to determine overall writing level (Norris and Ortega 2009). The dynamic and multidimensional realities of the CAFF constructs were thus measured in terms of their general and specific sub-constructs with intentions of eliminating unnecessary redundancy (see Norris and Ortega 2009). That is, some measurements of complexity, accuracy, fluency, and functionality, when used together, may be redundant 'because they tap the same measurable dimension of the construct and, conversely, there are other measures that are distinct and complementary' (2009:562). To avoid this redundancy, 11 measures were chosen for model development.

Eight of the 11 measures derive from Complexity Theory used to measure complexity, accuracy, and fluency (e.g., Biber, Gray and Poonpon 2011, 2013, Ortega 2003, Wolfe-Quintero, Inagaki and Kim 1998), and the remaining three measures from SFL (Halliday and Matthiessen 2014) for evaluating semiotic resources for constructing content, style and organization (Martin 2004, Martin and Rose 2003, Martin, Matthiessen and Painter 1997) (see Figure 1). For example, the measure for content was derived from the ideational metafunction of SFL that denotes the grammar of experience or experiential meaning. Frequency of verb types per clause is a measure to represent a portion of the transitivity system, which construes the world of human experiences into a manageable set of process, or verb, types. Process types may be material, mental, behavioral, verbal, relational, or existential (see Martin et al (1997:228) for further description). These processes represent inner and outer experiences of the world and are significant for showing how a learner constructs content.

All measures are based on frequency of linguistic occurrences due largely to the limitations in the computational techniques used to build the tool, but information from these measures can be used for further analysis to better represent the theoretical intentions. Algorithms could also be developed to align with external measures, such as the Common European Framework of Reference for Languages (CEFR) (see Shermis (2018) for a discussion about the CEFR and automated scoring).

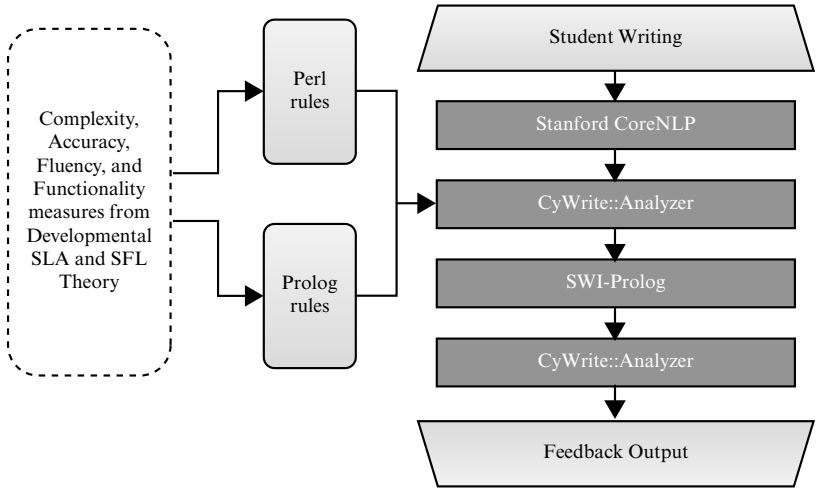
To construct the algorithms for multidimensional measures of CAFF, a hybrid approach to natural language processing, which included (1) statistical parsing of student texts and (2) rule-based feature detection to support score generation, was used. Figure 2 illustrates the simplified architecture of the hybrid approach to development. Details of the design process are described in Link (2015). In brief, in the first stage, the process of developing the algorithms starts with a student's text. Each sentence in a given text is parsed, or split into phrases or clauses, through Stanford CoreNLP (Manning and Schütze 1999). Stanford CoreNLP integrates a set of natural

Figure 1 CAFFite scoring model (measures represent complexity, accuracy, and fluency measures from Complexity Theory and functionality from Systemic Functional Linguistics)



language analysis tools (e.g., part-of-speech (POS) tagger and the parser) into a framework so that students' writing can be given linguistic annotations (e.g., clause-, phrase-, and word-level tags, also known as Stanford Typed Dependencies, see De Marneffe, MacCartney and Manning (2006)). These annotations are exported to Extensible Markup Language (XML) format for additional high-level text processing techniques.

Figure 2 Simplified architecture of the hybrid approach to scoring model development (adapted from Chukharev-Hudilainen and Saricaoglu 2014)



In the second stage of the hybrid approach, the parsed text is run through an analyzer fully adopted from Chukharev-Hudilainen and colleagues at Iowa State University (see Chukharev-Hudilainen and Saricaoglu 2014). This analyzer (CyWrite::Analyzer) processes the output from the parsed student text using Prolog logical programming language, which is rooted in first-order logic (see Bramer 2013). Use of the analyzer allows for separate programs (e.g., algorithms built for CAFFite) to be integrated into the architecture for various text analysis purposes. A subset of the CAFF features (clauses, phrases, appreciation resources, and conjunctions) were detected using Prolog algorithms. The remaining features were detected using Perl, another general purpose programming language.

Training of the scoring model was derived from regression-based analysis as a start-up method of predicting students' developmental level. Regression-based analyses have been used by Landauer, Laham and Foltz (2003) to determine the optimal set of features and weights for each of the features to best model the score of each essay. The first version of e-rater (Burstein et al 1998) used a stepwise regression technique to select the best features that are most predictive for a given set of data, and Project Essay Grade is also based on regression analysis (Page 1994). While it is worth noting that there are other approaches to machine learning, such as deep neural networks (Xi, Higgins, Zechner and Williamson 2012) and ranking supervised machine learning models (Yannakoudakis, Øistein, Geranpayeh,

Briscoe and Nicholls 2018), the advantage of the regression approach is that optimal solutions can be discovered for constructing a reliable score with respect to some measure of agreement between human and automated scores. This approach is limited, however, because the writing features may contribute to the regression equation differently depending on the kinds of information included (e.g., confounding variables that may or may not be included can result in different solutions in different applications) and the way the equation is used. In other words, a writing feature may contribute positively to the score in one model equation and negatively in another (Attali and Burstein 2006). These possibilities are common in practice because automated writing evaluation tools are often based on too many features to control and that overlook the serious problems that individual students may display.

With these concerns taken into consideration, regression analyses were run by using essays from a training set of 199 essays. Foltz, Rosenstein, Lochbaum and Davis (2012) found that performance on a scoring model increased as the training set size increased; results indicated that at about 200 essays in the training set the performance began to level off. Thus, 'the majority of the performance benefit can be obtained with 200 essays in the training set' (2012:5). This training contributed to the first iteration of the CAFFite scoring model that was further tested using human evaluation. The first iteration is reported in this chapter; further research and development of CAFFite is currently underway.

Human–human, human–computer reliability estimates: *A priori* and *a posteriori* evidence

Attali and Burstein (2006) argue that because computers do not understand writing as humans do, human evaluations should be a part of the validation process. This involvement requires development of rater guidelines and rater training as well as human–human reliability estimates across writing tasks and genres to set a benchmark for accuracy of the scoring model. Expert human raters should contribute both *a priori* and *a posteriori* validity evidence. Before the scoring model is developed, humans can be used to set a gold standard for the accuracy of the model and to determine relevance, or the expected performance of the assessment tool. After the model is developed, and even during ongoing development, human ratings can be compared to the computer's score for determining reliability and consistency of the tool. Calculations can be completed across raters, occasions of assessment, versions of writing tasks, and different levels of test-taker performance.

Prior to calculating reliability estimates, five expert human raters were extensively trained to identify the CAFF features in 20 training

essays (about 10% of the dataset) and then to holistically evaluate them using a holistic scoring rubric that reflects the 2012 ACTFL Proficiency Guidelines for Writing. Once reliability was established, a final coder analyzed all 199 essays for 11 measures over an extended timeframe to reduce coder fatigue. To evaluate the scoring engine's consistency across writing prompts, the same coder also analyzed an additional 84 essays collected from a different group of students responding to a different writing task prompt.

Codes were compared to the scoring engine's output to evaluate the performance of CAFFite in comparison to humans and across writing prompts. Correlation analysis was performed to determine human–human and human–computer consistency. Correlation measures the linear relationship between variables providing indicators of direction and strength between two variables (De Veaux, Velleman and Bock 2011). This method of analysis was used because the variables are continuous and not nominal like in the holistic ratings. Previous studies on automated scoring have frequently used correlation to investigate the reliability of computer-generated scores by correlating them with human scores (e.g., Burstein et al 1998, Page 1994).

Agreement statistics were calculated for comparison of holistic ratings. During the rating process, a minimum of two random raters evaluated each essay. On occasion a third (or sometimes a fourth) rater was asked to provide a final decision. Therefore, diagnostic assessment is determined using multiple raters and the same raters did not consistently rate the same set of essays. Agreement was thus determined using Krippendorff's alpha (α) (Hayes and Krippendorff 2007, Krippendorff 2011), which allows for the calculation of reliability 'regardless of the number of observers, levels of measurement, sample sizes, and presence or absence of missing data' (Hayes and Krippendorff 2007:77). In most automated scoring research, agreement is calculated using Cohen's kappa (Cohen 1968); however, because the ratings from humans are on a three-point scale and the ratings from CAFFite are dichotomous,[2] all assumptions are not met, and thus Cohen's kappa is inappropriate. Furthermore, Cohen's kappa does not take into account the degree of disagreement between raters and all disagreement is treated equally as total disagreement. Krippendorff's α is calculated so that different levels of agreement can contribute to the value of α.

The results that would provide the evidence necessary for supporting scoring validity are human–computer correlation coefficients that reach at

2 The dichotomous nature of the CAFFite score was chosen based on results from an unpublished pilot study that indicated low predictability in determining students' development based on a three-point scale. The scale is determined as pass or fail. A passing score means students do not have to take a second language writing course. A failing score means that they do.

least .70 and/or reach the level of acceptability in human score degradation, which means that human–computer reliability should not be more than .10 lower than the human–human reliability (Williamson, Xi and Breyer 2012). As shown in Table 2, the human coders were generally consistent among themselves when coding constituent features that contribute to the algorithms in the CAFFite scoring model. At the lower bound of significance was coding for phrases (Spearman's $\rho = .698$) at a .05 significance level. Coding of mental verbs (Spearman's $\rho = .611$) also did not reach the standard.

For holistic scoring, human–human inter-rater reliability for placement on an interval scale of 1–3 levels was interval $\alpha = .62$, which is borderline between 'fair' and 'good' (Strijbos and Stahl 2007). Agreement was the same when data were transformed to reflect the binary nominal variable of pass (Level 3) or fail (Levels 1 and 2): ordinal $\alpha = .62$. Using human final holistic ratings and the CAFFite holistic score, agreement was very poor, ordinal $\alpha = .09$, which means the computer's decision seemed to be random.

Although the agreement was lower than expected, it is not completely surprising that the automated score did not agree with the human ratings. First, human–human agreement was below an acceptable level for phrases and mental verb types. This marks the importance of ensuring rater training

Table 2 Spearman's rank order correlation for human–human and human–computer inter-coder reliability of constituent features of the CAFFite scoring model

CAFF feature	P	
	human–human[a]	**human–computer[b]**
T-units	.976**[c]	.918**
Dependent clauses	.756**	.742**
Phrases	.698*	.690*
Error-free T-units	.958**	.735**
Conjunctions	.773**	.295**
Appreciation	.755**	.425**
Verb types		
Material	.738**	.736**
Relational	.760**	.636**
Mental	.611**	.605**
Verbal	.885**	.554**
Existential	.877**	.792**
Behavioral	.875**	.916**

[a] *Coding was based on an N-size of 20 student texts.*
[b] *Coding was based on an N-size of 195 student texts. Four texts were excluded due to an unidentifiable server error.*
[c] ***Correlation is significant at the .01 level (2-tailed); * Correlation is significant at the .05 level (2-tailed).*

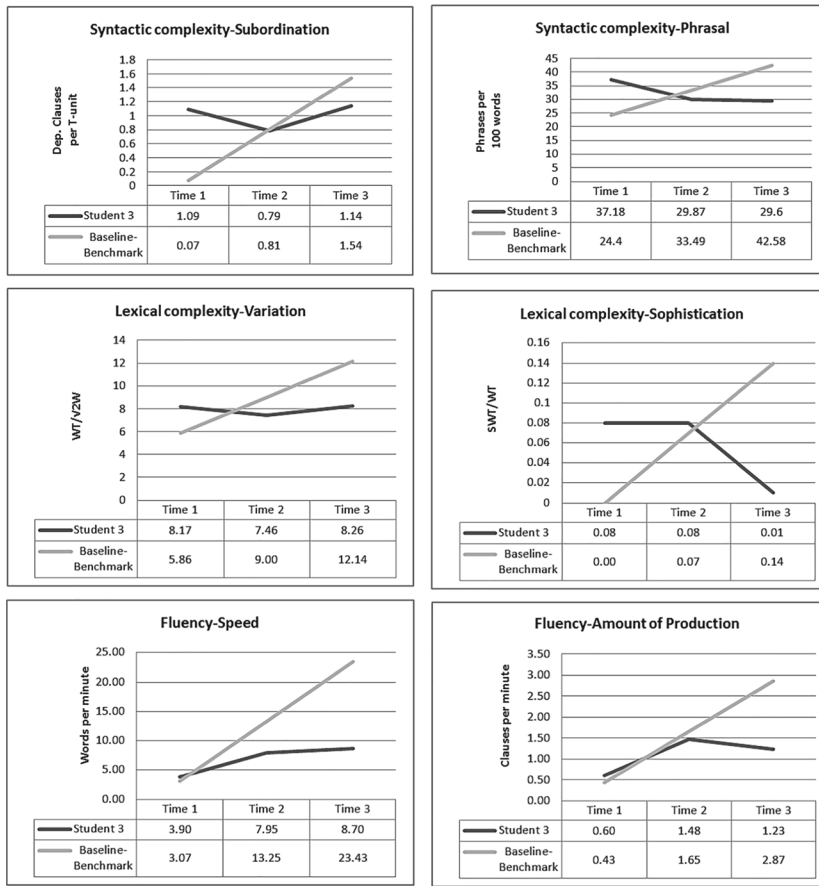
is sufficient because this low agreement largely affects the computer's ability to accurately predict students' developmental level since the computer is trained based on human judgments. If human judgments are not reliable, the computer would have difficulty producing a reliable score as well. In addition, the regression-based scoring model was not designed to take into account the several measures that could not be validated based on the human–computer reliability analyses (i.e., dependent clauses, error-free T-units, appreciation resources, verbal processes, and conjunctions). The low accuracy and reliability in the detection of these features could have contributed to the low reliability in producing a holistic score. Most importantly, however, is that CAFF measures may not be able to predict students' overall level of performance due to the nonlinear development of individual sub-constructs. That is, there may be tradeoff effects between the sub-constructs because of task effects (Robinson 2005), such as task complexity (cognitive factors), task conditions (interactional factors), and task difficulty (learner factors). Each of these could impact the reliability of automated scores from student to student and from context to context. Larsen-Freeman and Cameron (2008:16) argue:

> Whereas positivist research is based on the assumption that there are universal laws and thus sets predictability as a goal of the research process, from this complexity theory perspective, no two situations can be similar enough to produce the same behavior; thus predictability becomes impossible.

The value of the holistic scores is therefore limited, but since the scoring model was developed with SLA theory in mind, the individual measures that are accurately identified can still shed light on students' developmental trajectories. Take the student in Figure 3 as an example. Student 3's developmental trajectory is shown here because he was labeled as an average-performing student based on writing performance at Time 1, 2, and 3, which are representative of the target use context (i.e., Levels 1, 2, and 3 as described in the section 'Target use situation and test-taker characteristics'). Baseline and benchmark scores are intended to be a comparison of the individual student to his peers in the same learning context rather than using native speaker samples as a gold standard. The scores were determined using the lowest and highest scores of each measure from the 199 essays in the training set of the *a priori* evidence illustrated earlier.

The student started Time 1 at a more advanced level of syntactic and lexical complexity compared to many of his peers but with low fluency. Interestingly, the student's fluency developed the furthest from Time 1 to Time 2 while other measures either decreased or stayed the same. From Time 2 to Time 3, lexical variation seemed to show the greatest improvement. At a

Figure 3 Example of CAFFite measures used to reveal an individual student's developmental trajectory



closer look, Student 3 used the word 'interest' six times in the essay at Time 2, as shown in the following examples:

g. In modern life, people consider their personal interest, intellectual interest, social honor, mon
le consider their personal interest, intellectual interest, social honor, money, family environment,
ng time. Therefore, in my opinion, their personal interest is the most important factor to choose th
was 19 years old. It says that the personal interest is one of the most important factor to
live happy, they have to work in their interest part to avoid boring life. It someone doe
life. If someone does not consider their personal interest the person will be not happy in life.

It is difficult, however, to attribute this word repetition to a lack of development in lexical variation because the essay prompts asked the student how personal, intellectual, or political interests affect career decisions:

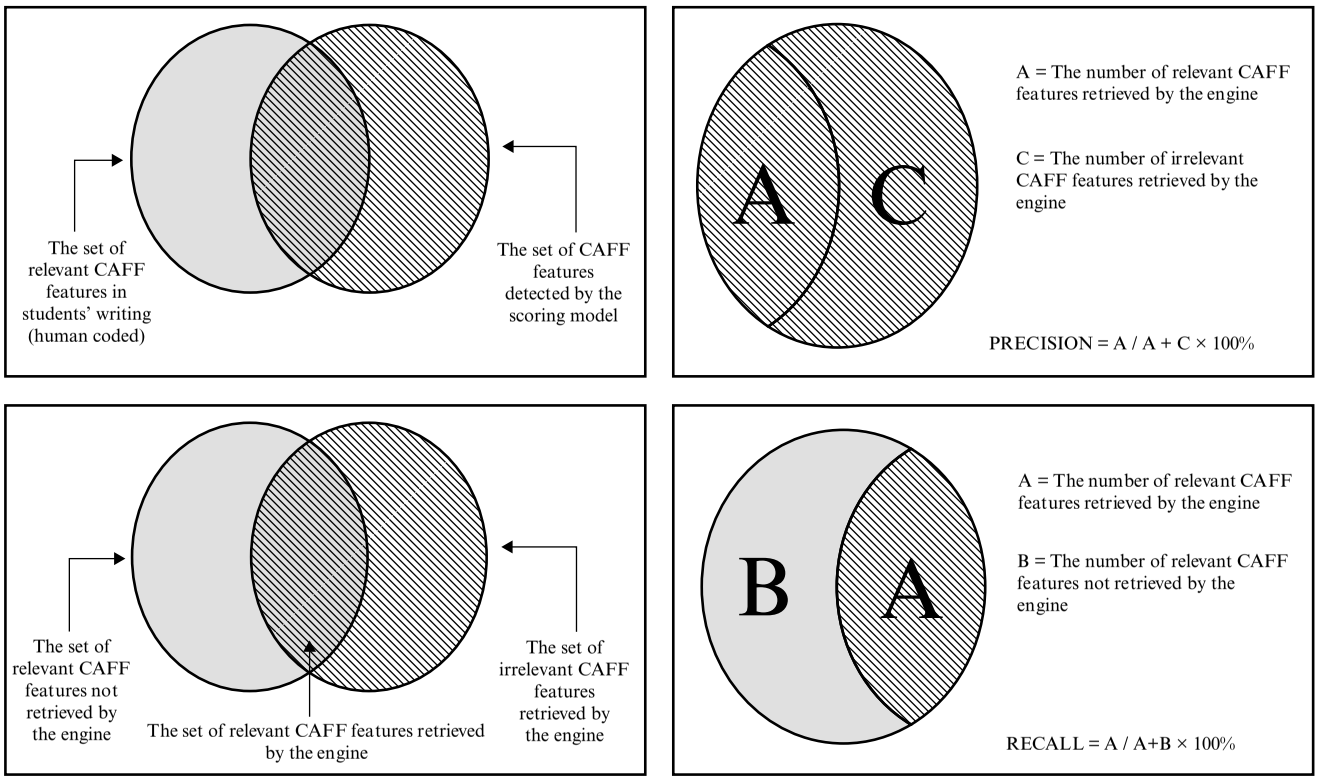
> In your opinion, what factors should a person consider when choosing a career? Should people follow their personal, intellectual, or political interests? Is salary the most important factor? What about status or family influence? Discuss which factors you think are most important for an individual to consider when choosing a career. Support your position with reasons and/or examples from your own experiences, observations, or reading.

This finding demonstrates the possible effects that prompts can have in the interpretation of scores, suggesting that adequate score interpretation materials should accompany the assessment and provide reference to the prompt used. Evidence of a prompt effect does not, however, go against the claim that the automated measures can be used to diagnose students' development. Instead the task or research design would have to be carefully constructed. Time-series modeling and analysis would be a next step in researching and validating the measures at an individual level. In this method of analysis, data from a series of data points over time are used to determine causal effects on a variable or change in a variable over time (Imdad Ullah 2013). These data points should be collected from tasks of similar nature (e.g., complexity, genre, time on task) so that an evaluation of a student's developmental path can be more accurate. Despite the potential for reflecting students' developmental trajectories, there is still a need to further validate that the measures are accurately applied into the feedback generation model in order to ensure appropriate interpretation of feedback. This accuracy can be evaluated using precision/recall and linguistic analysis.

Precision/recall and linguistic analysis: *A posteriori* evidence

Precision and recall are measures that assume a set of features in a student's text are relevant to the measures being captured by the system's engine (Barnhart, Haber and Lin 2007). Precision is the ratio of the number of relevant writing features retrieved to the total number of irrelevant and relevant features retrieved (usually expressed as a percentage). See the top of Figure 4 for how precision was calculated for CAFFite. Recall is the ratio of the number of relevant writing features the engine retrieved to the total number of relevant features a human coder retrieved, also expressed as a percentage (Cowie and Wilks 2000, Jackson and Moulinier 2007, Manning and Schütze 1999). Recall is further illustrated at the bottom of Figure 4. More formally, precision (or confidence) and recall (or sensitivity) are calculated based on true and false positives (TP/FP), referring to the number of predicted positives that were correct/incorrect, and true and false negatives (TN/FN), the number of predicted negatives that were correct/incorrect. The F1-score accounts for both the false positives and the false negatives.

Figure 4 Precision and recall of relevant and irrelevant CAFF features in students' writing



This score is a calculated mean, or weighted average, of precision and recall defined by the following equation:

$$F1 = 2 \times (\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$$

Quinlan, Higgins and Wolff (2009) suggested reaching 80% precision before integrating a feature into a scoring model. The higher the precision indicates the engine performs similarly to human coding with possible tradeoff effects between precision and recall. Table 3 shows results from the first iteration of CAFFite score modeling. The engine performed well in detecting eight CAFF constituents (T-units, phrases, conjunctions, and material, relational, mental, existential, and behavioral verbs), demonstrating levels of precision well above the 80% standard. However, four constituents were detected with low accuracy. Dependent clauses were detected with precision = .69, recall = 1, F1 = .82. Error-free T-units were detected with precision = .64, recall = 1, F1 = .78. Appreciation resources were detected with precision = .59, recall = 1, F1 = .74. Verbal verbs were detected with precision = .67, recall = 1, F1 = .80.

Precision, recall, and F1-score analysis have their advantages, which is why they are commonly used techniques in automated scoring research, but they are nonetheless still biased measures. Precision, recall, and F1-score analyses ignore performance in correctly handling negatively extracted language features and fail to take account of the chance-level performance (Powers 2011). Therefore, when low precision was detected, linguistic analysis of a subset of essays (n = 20) was conducted to complement the quantitative findings and provide added evidence in appraising the scoring validity of the CAFFite engine. Table 4 shows the sources of error in detecting three of the four CAFF features with low precision. Error-free T-units were excluded from the analysis because the known source of error is due to the use of open source proofreading software as an external reference for error detection (see www.languagetool.org). The types of errors identified by LanguageTool (LT) varied from those identified by humans who used the guidelines in Polio (1997). For example, LT counts spelling errors whereas the humans did not. Therefore, evidence cannot be provided in support of the use of the error-free T-units/T-units measure for gauging students' development in accuracy.

The second column in the table identifies the total number of features detected by CAFFite. The next column shows how many of the total features detected were irrelevant, meaning that the computer detected a feature that the human coder did not. The last set of columns indicates the source of the errors.

For example, appreciation resources (54%) were inaccurately detected primarily due to CAFFite rule issues. This was mainly because of multi-word

Table 3 Precision, recall, and F1-score of the CAFFite scoring engine with human codes as standard reference (N = 86[a])

CAFF feature	Total features detected by human coder	Performance of the CAFFite scoring engine					
		Relevant features retrieved	Irrelevant features retrieved	Relevant features not retrieved	Precision	Recall	F1-score
T-units	1,737	1,737	167	0	0.91	1.00	0.95
Dependent clauses	734	734	329	0	0.69	1.00	0.82
Phrases	5,666	5,666	1,508	0	0.79	1.00	0.88
Error-free T-units	936	936	522	0	0.64	1.00	0.78
Conjunctions	1,679	1,589	0	90	1.00	0.95	0.97
Appreciation resources	2,358	2,358	1,652	0	0.59	1.00	0.74
Verb types							
Material	1,251	1,251	234	0	0.84	1.00	0.91
Relational	707	707	104	0	0.87	1.00	0.93
Mental	333	333	18	0	0.95	1.00	0.97
Verbal	117	117	59	0	0.67	1.00	0.80
Existential	68	68	0	0	1.00	1.00	1.00
Behavioral	107	107	27	0	0.80	1.00	0.89

[a] *Four texts were excluded due to an unidentifiable server error.*

Table 4 Source of errors in feature detection of the CAFFite engine (N = 19)[a]

CAFF feature	Total features detected	Irrelevant features retrieved	Source of errors							
			CAFFite rule issues		Learner language problems		Syntactic parser failures		Unclassified	
			N	%	N	%	N	%	N	%
Dependent clauses	236	73	5	6.8	44	60.3	9	12.3	15	20.5
Error-free T-units										
Appreciation resources	891	367	198	54.0	55	15.0	38	10.4	76	20.7
Verbal processes	39	13	10	77.0	3	23.1	1	7.7	0	0.0

[a] *One essay was excluded due to a server error.*

expressions, transition words, and subordinating conjunctions tagged as adverbs. Below are a few examples with irrelevant features underlined:

(1) Our live become more and more easy.
(2) Moreover, modern conveniences freed people for doing the same staff again and again.
(3) As far as I am concerned, I strongly believe that modern conveniences are beneficial.

In the first two examples, 'more and more' and 'again and again' were manually coded as one appreciation resource each whereas CAFFite counted two resources. Transition words and subordinating conjunctions tagged as adverbs were also problematic. Words like 'however', 'moreover', 'then', 'first' and 'second' were all automatically tagged as appreciation resources but not manually coded. In the third example, the word 'far' was not manually coded as a resource but was counted by CAFFite because the parser tagged the word as an adverb while the human coder recognized it as a set expression and thus did not count it.

As another example, learner language problems were the most common source of error in the identification of dependent clauses (60.3%). Errors due to language use were verified by correcting the students' errors and then running the sentence through CAFFite a second time. If the irrelevant clause was no longer detected, the source of the error was coded as a learner language problem. Any type of grammatical, word usage, or mechanical error seemed to affect the accuracy of the parser and thus the precision of CAFFite's detection of dependent clauses. The sentences below provide examples of each type of error. The underlined parts of each sentence indicate an irrelevant dependent clause that CAFFite identified. The first sentence is a fragment, a grammatical error. The second sentence has a wrong word usage error; the word 'convince' should be in the adjective form not the verb form. The third sentence has an error in mechanics. Here, the student wrote a run-on sentence. When the comma was omitted, the parser recognized the final independent clause as a dependent clause. When the comma was added after the word 'money', the parser was able to correctly parse the sentence so that no dependent clause was detected. This demonstrates that the error was truly due to the learner's language and not the parser.

(1) For example, not only using them for calling, playing games, search on the internet, see the email and so on.
(2) Second, in the city, there are many convince machines such as the train station, bus and taxi system.
(3) Modern convenience helped us gain life and money so it must be important for our life.

Ongoing linguistic analysis is possible until a feature reaches the gold standard in precision (at least 80%). The efficiency of the algorithm(s) can then be enhanced by either refining the logic of the measures chosen or by altering the data input to train the engine (Kowalski 1979).

Evaluation of scoring validity

Once the necessary *a priori* evidence and *a posteriori* evidence is collected, a more rigorous evaluation of scoring validity can be made. It should be noted that different pieces of evidence from those described here may be collected as long as the data are answering the critical question underlying the validation component. That is, 'How far can we depend on the scores which result from automated assessment?' In this chapter, I introduced how to utilize Shaw and Weir's (2007) socio-cognitive framework to evaluate scoring validity for diagnostic writing assessment systems. I then demonstrated how the CAFFite feedback engine was trained to extract features from essays using validated measures of quality writing performance (*a priori* evidence) to provide human–human (*a priori* evidence) and human–computer (*a posteriori* evidence) reliability estimates that are consistent with different performance levels. However, further evidence should be collected to determine reliability over occasions of assessment and versions of writing tasks to provide adequate and appropriate diagnostic feedback. I was also able to demonstrate the use of precision/recall and linguistic analyses (*a posteriori* evidence) that showed accurate application of eight measures into the feedback scoring model. Additional linguistic analyses would need to be conducted to validate the implementation of additional features.

Overall validity judgment of the automated evaluation engine should not be made until evidence from each validation component (as outlined in Table 1) has been collected and analyzed. In this way, validity is viewed as a unitary concept. The next step is to communicate these results to stakeholder groups who may not be versed in jargon relevant to automated feedback development and evaluation. The prospect of providing individualized diagnostic feedback on SLA-informed measures of writing development can help account for the educational culture and social context of language use in which assessments are to be implemented. This added attention can provide interesting information for future development of engines that are context-specific and aimed at a target language use while limiting the over-generalization of test-takers as holistic, linear beings rather than ever-changing, dynamic individuals. One pathway to completing such development is through the lens of language acquisition theory, such as the attempt to develop CAFFite using two theories that see language as a system: Complexity Theory and SFL. By bringing together developers and language testers informed by SLA, a set of theoretically and empirically based standards and measurements can be

developed to frame automated feedback generation. Automated output can then provide not only a holistic score based on validated measures but also individual scores on constructs such as complexity, accuracy, fluency, and functionality that can provide an in-depth view of students' writing profiles and developmental trajectories when evaluated over time.

For efforts to succeed, however, engines must meet the needs and expectations of one set of key stakeholders: language instructors. Language teachers are in many ways the gatekeepers for technology integration in the classroom. If convinced of the legitimacy of automated feedback, computer-based diagnostic assessment may find a regular place in the classroom assessment context for evaluating language development, the complex nature of the writing process, and the role of automated writing evaluation in shaping social and cognitive processing.

Conclusion

Technology has transformed how automated writing assessment operates. Its widespread implementation in high-stakes testing has stakeholders questioning the validity of automated feedback use. Yet, with the potential for diagnostic assessment, it is without much debate that automated feedback will only continue to advance. Thus, discussions must begin focusing on how the power and efficiency of new models can reshape writing analysis for broader intentions, such as stakeholder use in educational and alternative assessment contexts. Simply repurposing existing engines continues to disregard specific target use situations and key user characteristics that influence how the scores are used. Furthermore, without input from specialists in language acquisition theory, measures chosen to develop algorithms or fine-tune large language models risk not catering to the needs of language practitioners and the understanding of individual student development. Instead, scoring models will continue to be centered on computation and programming using complicated scoring models that do little to inform theory-building and empirical research within second language writing development.

As language assessment moves forward in the digital age, knowing more about what is in the black box of automated feedback generation should add to stakeholders' trust in automated score use. Moreover, clear reporting of validity evidence using existing validation frameworks can help to structure an argument for or against diagnostic assessment systems. This evidence, however, should not be disguised. As shown with CAFFite, not all evidence was positive. Precision at the very onset of linguistic analysis had a negative impact on score validity. This transparency helps to document the progress and evolution of a model and provide a springboard for future development.

Frameworks, such as Shaw and Weir's (2007) socio-cognitive approach, can also determine how current deployments of automated feedback engines can be

systematically evaluated for language assessment for and beyond high-stakes testing. The possibilities of using automated scoring for measuring, collecting, analysing, and reporting data about learners and their contexts seems to have a bright future for optimizing learning and teaching. Validity evidence summarized in this chapter can provide strong pedagogical implications of wider significance by suggesting what evidence is needed to bring the focus of automated writing evaluation on key stakeholders, thus assisting in an understanding of their individual social, cognitive, and linguistic needs.

References

ACTFL (2012) *ACTFL Proficiency Guidelines*, available online: www.actfl.org/uploads/files/general/ACTFLProficiencyGuidelines2012.pdf

ACTFL (2017) *World Readiness Standards for Learning Languages*, available online: www.actfl.org/sites/default/files/publications/standards/World-ReadinessStandardsforLearningLanguages.pdf

Attali, Y and Burstein, J (2006) Automated essay scoring with e-rater V.2, *Journal of Technology, Learning and Assessment* 4 (3), 1–21.

Attali, Y, Bridgeman, B and Trapani, C (2010) Performance of a generic approach in automated essay scoring, *Journal of Technology, Learning, and Assessment* 10 (3), available online: ejournals.bc.edu/index.php/jtla/article/view/1603

Barnhart, H, Haber, M and Lin, L (2007) An overview of assessing agreement with continuous measurement, *Journal of Biopharmaceutical Statistics* 17 (4), 529–569.

Biber, D, Gray, B and Poonpon, K (2011) Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?, *TESOL Quarterly* 45 (1), 5–34.

Biber, D, Gray, B and Poonpon, K (2013) Pay attention to the phrasal structures: Going beyond T-units – A response to WeiWei Yang, *TESOL Quarterly* 47 (1), 192–201.

Bramer, M (2013) *Logic Programming with Prolog*, London: Springer.

Burstein, J and Chodorow, M (2003) Directions in automated essay scoring, in Kaplan, R (Ed) *The Oxford Handbook of Applied Linguistics*, Oxford: Oxford University Press, 487–497.

Burstein, J and Chodorow, M (2010) Progress and new directions in technology for automated essay evaluation, in Kaplan, R (Ed) *The Oxford Handbook of Applied Linguistics* (Second edition), Oxford: Oxford University Press, 487–497.

Burstein, J, Kukich, K, Wolff, S, Lu, C, Chodorow, M, Braden-Harder, L and Harris M D (1998) Automated scoring using a hybrid feature identification technique, in *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics*, East Stroudsburg: Association for Computational Linguistics, 206–210.

Chapelle, C (2001) *Computer Applications in Second Language Acquisition: Foundations for Teaching, Testing and Research*, Cambridge: Cambridge University Press.

Chapelle, C (2009) The relationship between second language acquisition theory and computer-assisted language learning, *The Modern Language Journal* 93, 741–753.

Chukharev-Hudilainen, E and Saricaoglu, A (2014) Causal discourse analyzer: Improving automated feedback on academic ESL writing, *Computer-Assisted Language Learning* 29 (3), 496–516.

Cohen, J (1968) Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit, *Psychological Bulletin* 70 (4), 213–220.

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

Cowie, J and Wilks, Y (2000) Information extraction, in Dale, R, Moisl, H and Somers, H (Eds) *Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*, New York: Marcel Dekker Inc, 241–260.

De Marneffe, M C, MacCartney, B and Manning, C D (2006) Generating typed dependency parses from phrase structure parses, in Calzolari, N, Choukri, K, Gangemi, A, Maegaard, B, Mariani, J, Odjik, J and Tapias, D (Eds) *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC '06)*, Genoa: ELRA, 449–454.

De Veaux, R D, Velleman, P F and Bock, D E (2011) *Stats: Data and models* (Third edition), Upper Saddle River: Pearson.

Deane, P (2013) On the relation between automated essay scoring and modern views of the writing construct, *Assessing Writing* 18 (1), 7–24.

Ferris, D R (2011) *Treatment of Error in Second Language Student Writing* (Second edition), Ann Arbor: The University of Michigan Press.

Foltz, P, Rosenstein, M, Lochbaum, K and Davis, L (2012) *Improving Reliability Throughout the Automated Scoring Development Process*, Vancouver: National Council on Measurement in Education.

Halliday, M A K (Ed) (1979) *Working Conference on Language in Education: Report to Participants*, Sydney: Extension programme and department of linguistics, Sydney University.

Halliday, M A K and Hasan, R (1989) *Language, Context, and Text: Aspects of Language in a Social-semiotic Perspective*, Oxford: Oxford University Press.

Halliday, M A K and Matthiessen, C M I M (2014) *An Introduction to Functional Grammar* (Fourth edition), New York: Routledge.

Hayes, A F and Krippendorff, K (2007) Answering the call for a standard reliability measure for coding data, *Communication Methods and Measures* 1 (1), 77–89.

Imdad Ullah, M (2013) *Time series analysis and forecasting*, Basic Statistics and Data Analysis, available online: itfeature.com/time-series-analysis-andforecasting/time-series-analysis-forecasting

Jackson, P and Moulinier, I (2007) *Natural Language Processing for Online Applications: Textual Retrieval, Extraction, and Categorization*, Philadelphia: John Benjamins Publishing Company.

Kowalski, R (1979) Algorithm = Logic + Control, *Communications of the ACM* 22 (7), 424–436.

Krippendorff, K (2011) Computing Krippendorff's alpha reliability, *Departmental Papers (ASC)* 43, 1–10.

Landauer, T K, Laham, D and Foltz, P W (2003) Automated scoring and annotation of essays with the Intelligent Essay Assessor, in Shermis, M D and Burstein, J (Eds) *Automated Essay Scoring: A Cross-disciplinary Perspective*, Mahwah: Lawrence Erlbaum Associates, 87–112.

Larsen-Freeman, D (2006) Functional grammar: On the value and limitations of dependability, inference, and generalizability, in Chalhoub-Deville, M, Chapelle, C and Duff, P (Eds) *Inference and Generalizability in Applied Linguistics*, Amsterdam: John Benjamins Publishing Company, 115–133.

Larsen-Freeman, D and Cameron, L (2008) *Complex Systems and Applied Linguistics*, Oxford: Oxford University Press.

Li, Z, Link, S, Ma, H, Yang, H and Hegelheimer, V (2014) The role of automated writing evaluation holistic scores in the ESL classroom, *System* 44 (1), 66–78.

Link, S (2015) *Development and validation of an automated essay scoring engine to assess students' development across program levels*, doctoral dissertation, Iowa State University.

Manning, C and Schütze, H (1999) *Foundations of Statistical Natural Language Processing*, Cambridge: MIT Press.

Martin, J R (2004) Mourning: How we get aligned, *Discourse & Society* 15, 321–344.

Martin, J R and Rose, D (2003) *Working with Discourse: Meaning Beyond the Clause*, London: Continuum.

Martin, J R, Matthiessen, C M I M and Painter, C (1997) *Working with Functional Grammar*, London: Edward Arnold.

Norris, J M and Ortega, L (2009) Towards an organic approach to investigating CAF in instructed SLA: The case of complexity, *Applied Linguistics* 30 (4), 555–578.

Ortega, L (2003) Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing, *Applied Linguistics* 24, 492–518.

Page, E B (1994) Computer grading of student prose, using modern concepts and software, *Journal of Experimental Education* 62, 127–142.

Polio, C (1997) Measures of linguistic accuracy in second language writing research, *Language Learning* 47, 101–143.

Powers, D M W (2011) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, *Journal of Machine Learning Technologies* 2 (1), 37–63.

Quinlan, T, Higgins, D and Wolff, S (2009) *Evaluating the Construct-Coverage of the e-rater® Scoring Engine*, ETS Research Report, available online: files.eric.ed.gov/fulltext/ED505571.pdf

Robinson, P (2005) Cognitive complexity and task sequencing: Studies in a componential framework for second language task design, *International Review of Applied Linguistics* 45, 1–32.

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

Shermis, M D (2018) Establishing a crosswalk between the Common European Framework for Languages (CEFR) and writing domains scored by automated essay scoring, *Applied Measurement in Education* 31 (3), 177–190.

Strijbos, J W and Stahl, G (2007) Methodological issues in developing a multi-dimensional coding procedure for small-group chat communication, *Learning and Instruction* 17 (4), 394–404.

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

Williamson, D M, Xi, X and Breyer, F J (2012) A framework for evaluation and use of automated scoring, *Educational Measurement: Issues and Practice* 31 (1), 2–13.

Wolfe-Quintero, K, Inagaki, S and Kim, H (1998) *Second Language Development in Writing: Measures of Fluency, Accuracy and Complexity*, Honolulu: Second Language Teaching and Curriculum Center, University of Hawai'i at Manoa.

Xi, X (2010) Automated scoring and feedback systems: Where are we and where are we heading?, *Language Testing* 27 (3), 291–300.

Xi, X, Higgins, D, Zechner, K and Williamson, D (2012) A comparison of two scoring methods for an automated speech scoring system, *Language Testing* 29 (3), 371–394.

Yannakoudakis, H, Øistein, E A, Geranpayeh, A, Briscoe, T and Nicholls, D (2018) Developing an automated writing placement system for ESL learners, *Applied Measurement in Education* 31 (3), 251–267.

12 Towards a validity argument for genre-based automated writing evaluation

Elena Cotos
Iowa State University, USA

Abstract

The uses of automated writing evaluation (AWE) have been increasingly investigated through the lens of the validity argument framework. It has been rightfully argued that validation requires complementary and alternative sources of evidence (Chung and Baker 2003, Williamson, Xi and Breyer 2012). Foregrounded by core theoretical tenets (Kane 2013), this study examined the evaluation inference in the interpretive argument (Chapelle 2012) outlined for a genre-based AWE system. The argument was consolidated with premises from the socio-cognitive validity framework (Weir 2005) such that both human and automated rhetorical analysis data, as well as the construct, take centre stage. The purpose was to investigate rebuttal evidence that weakens the assumption that the rhetorical feedback generated by this system's engine accurately reflects learners' genre writing ability. Quantitative measures (percent agreement, Cohen's kappa, confusion matrixes) along with qualitative analyses were employed to: a) determine degrees of reliability and understand differences affecting human–AWE agreement, and b) detect human and AWE performance errors and understand underlying reasons. The results suggest that reliability was affected by a fundamental limitation of the engine's mono-label approach. With regard to performance, the engine exhibited error patterns similar to humans, revealing that functional language features may be the key sources of feedback inaccuracy. Implications extend beyond the evaluation inference, as this work provides a discerning picture of the complexity and difficulty of automated rhetorical analysis, and contributes to the nascent genre-based AWE field of practice with two new heuristic taxonomies for rhetorically driven analysis of interpretive judgment (human input) and automated detection (computer output) of communicative intent.

Introduction

With the advent of innovative instructional technologies, the relationship among teaching, assessment, and learning has been fostered in a wide variety of contexts. Writing instruction, in particular, has witnessed large-scale uses of automated writing evaluation (AWE) systems such as Criterion (Educational Testing Service), WriteToLearn (Pearson Education), MyAccess! (Vantage Learning), Revision Assistant (Turnitin), Folio (Measurement Inc.), etc. AWE systems automatically analyze students' writing and provide formative feedback on different writing traits (e.g., lexico-grammatical errors, topic relevance, discourse structure, genre conventions, style, mechanics, etc.). The automatically generated feedback can be used to prioritize learning and improvement of different aspects of the writing construct, potentially serving the purpose of learning-oriented assessment (Jones and Saville 2016, Purpura and Turner 2014) for writing development.

The uses of AWE are increasingly evaluated through the lens of the validity argument framework (Kane 2011), which has been applied to both high-stakes and low-stakes language assessments, including formative and learning-oriented (e.g., Gleason 2013, Gruba, Cárdenas-Claros, Suvorov and Rick 2016, Link and Li (Eds) 2018, Yang and Cotos 2018). Multi-faceted evidence has been reported to demonstrate whether and to what extent certain inferences about the use of AWE feedback could be supported (Chapelle, Cotos and Lee 2015, Li, Link, Ma, Yang and Hegelheimer 2014, Ranalli, Link and Chukharev-Hudilainen 2017).

The evidence accumulated for those inferences, however, have yet to be consolidated into a coherent validity argument, for individual tools and then for AWE in general. This is a sizable challenge due to the myriad contexts in which AWE systems are used and the various forms of feedback they afford. Plus, the design of AWE technology itself has not been without controversy, in part because the writing construct was defined differently in terms of automated scoring goals as opposed to pedagogical goals (Cotos 2015). As Deane (2013:12) explains, 'this disjuncture is due to the contrast between models focused on "text quality", measured in the end product, versus models focused on "writing skill", which is an attribute of the writer, not the text'. This construct-related issue is unequivocally interconnected with validity concerns.

This chapter considers both construct definition and validity with reference to an exemplar of rhetorically driven, genre-based AWE – the Research Writing Tutor (RWT) (Cotos 2016, vimeo.com/90669213). Validation is pursued in order to be confident that the feedback is appropriate for the intended purpose of developing genre writing competence, and that appropriate pedagogical and learning decisions are made based on this form of AWE assessment. The construct definition adopted in lieu of pedagogical purposes is that genre writing competence is a composite of the writer's metacognitive knowledge

of the rhetorical task, socio-disciplinary awareness of the target disciplinary community, and metapragmatic ability to produce a written artifact as a communicative action realized with rhetorically appropriate language choices (Cotos 2014). This definition resonates with the more general view that the writing construct includes 'the rhetorical ability to integrate an understanding of audience, context, and purpose when both writing and reading texts' and 'the ability to learn and use the conventions appropriate to a specific genre of writing' (Perelman 2012:129).

Foregrounded by core theoretical tenets, the validation approach builds on a pre-evaluation interpretive argument for RWT with a series of inferences and assumptions, which guide the collection of systematic evidence needed to ultimately build a validity argument in support of the uses of its feedback. For more explicit connections between empirical evidence and specific types of validity, the interpretive argument integrates premises from the socio-cognitive validity framework (Weir 2005) where construct validity is treated as a symbiotic relationship among context, scoring, and cognitive validity.

Presenting validity evidence geared towards the evaluation inference in the argument, the second part of the chapter elaborates on an exploratory investigation of the accuracy of the feedback engine. Related to the conception of scoring validity in Weir's framework, and somewhat expanding it, is the construct relevance of the computational models and the text features from which the automated feedback is derived. Unlike traditional AWE systems that leverage the capabilities of automated scoring, RWT employs a different approach to automated text analysis and feedback. Therefore, supporting the claim that RWT's feedback accurately targets relevant areas for revision, improvement and learning requires more than conventional evidence of reliability, especially because the evidence accumulated so far partially supports the assumption that RWT's analysis engine can accurately detect important rhetorical traits of the target genre. The exploration thus aimed to understand what might affect accuracy. Quantitative measures and qualitative analyses of automated output vis-à-vis human annotation were employed to examine both reliability and performance errors in identifying rhetorical traits. The results generated unique observations and valuable heuristic tools, which are recommended for successive study and improvement of RWT feedback accuracy, as well as for the study of scoring validity of future genre-based AWE tools.

Research Writing Tutor (RWT) – a genre-based AWE exemplar

RWT is a tool for genre writing and revision used for instruction and for self-paced learning by L1 and L2 writers in English for academic purposes contexts such as graduate writing courses, writing workshops, and

one-on-one writing consultations. The target genre is the research article (RA), and more specifically the Introduction, Methods, Results, and Discussion/Conclusion (IMRD/C) sections. Central to the design of RWT and its feedback is the construct-related idea that genre writing competence entails rhetorical knowledge of the genre's intended purposes and formal knowledge of the textual instantiations (e.g., structure and discourse form) of the genre (Tardy 2009). RWT attends to these aspects of the genre writing construct by operationalizing the core tenets of genre theory (Swales 1981, 1990) – rhetorical moves (communicative goals) and steps (functional strategies) – that are characteristic of the target genre. Practically, it reproduces the rhetorical conventions of IMRD/C writing in 30 disciplines through a range of affordances, from video explanations of individual moves and steps in the Learning Module, to concordance examples in the corpus-based Demonstration Module, and to various rhetorical prompts generated by the Feedback Module (see Cotos 2016). The Feedback Module provides different types of feedback, briefly described and exemplified in Figure 1.

The feedback is generated by RWT's AWE engine, which runs a suite of machine learning models that automatically classify every sentence in students' drafts into rhetorical traits (i.e., moves and steps). The classification models were trained using a corpus of authentic RAs published in the top journals of those disciplines. The corpus contains IMRD/C sub-corpora that were manually annotated with empirically derived move/step traits for each section. This annotated text data allowed for training classifiers, or computational models, to detect the moves and steps that are specific to individual IMRD/C section texts. For instance, the Introduction classifiers classify individual sentences into one of the three moves and one of their respective steps based on Swales' (1981, 1990) Create-A-Research-Space model presented in Figure 2.

More specifically, the move of a sentence is first predicted by the move classifier; then, the step of the sentence within that move is predicted by the step classifier (see Cotos and Pendar 2016). This way, the move classification is translated to color-coded feedback on each sentence (blue for Move 1, red for Move 2, and green for Move 3), and the step classification is rendered in the functional feedback (e.g., 'It looks like you are claiming centrality.'). The numerical and goal-orienting types of feedback compare the move and step classification outcomes with the distribution of rhetorical traits in students' target disciplines, which are represented by the above-mentioned corpus.

Towards a validity argument for genre-based AWE

Validity argument framework

Validation research has evolved through a number of approaches adopted in language testing, with the argument-based approach gaining most

Figure 1 RWT feedback

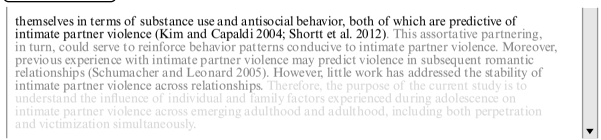
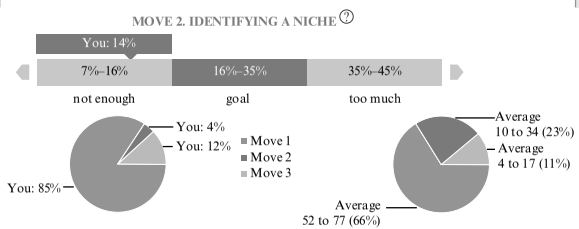
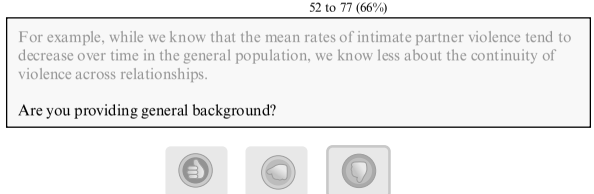
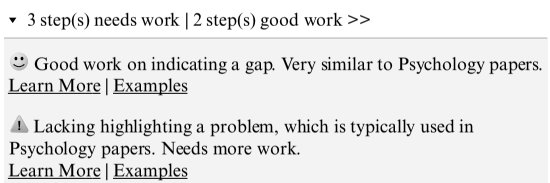
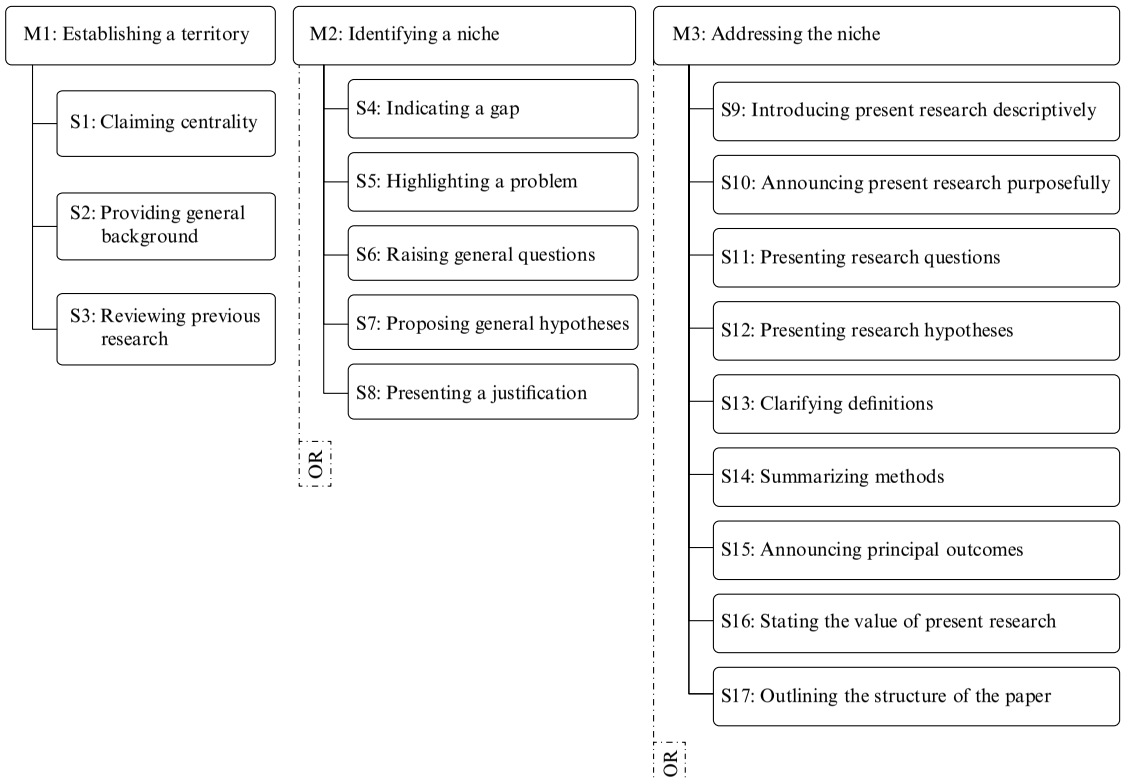
Type of feedback	Emphasis	Example
Color-coded (entire text is shown in colors representing particular moves)	Textual context (depicts the move structure of the draft)	
Numerical (range bars and pie charts summarize the distribution of moves in the draft)	Socio-disciplinary context (compares the move distribution in the draft with the target discipline)	
Functional (comments or clarifying questions about the step function of individual sentences)	Monitoring of rhetorical intent (draws attention to functional meaning)	
Goal-orienting (the range bars for moves expand to show which step/s are addressed well and which may be lacking or needing more work)	Planning and revision (sets goals based on textual and socio-disciplinary context)	

Figure 2 Rhetorical traits used to classify sentences in Introduction texts



popularity in the past few decades (Chapelle and Voss 2014). This approach builds on the notion of validity argument, where argument refers to a chain of reasoning that forms a structured rationale. Kane (2011, 2013) presents the validity argument as a framework essentially consisting of two components: 1) interpretive and use argument – a detailed account of proposed interpretations and uses of assessment results, and 2) validity argument – an empirical evaluation of the plausibility of the interpretive argument. The interpretive argument outlines specific inferences starting with the characteristics of observed performances and test tasks (grounds) and leading up to conclusions based on the assessment (claims). Inferences have to be justified by warrants, i.e., generally held rule-of-thumb principles (Chapelle, Enright and Jamieson (Eds) 2008). Warrants rely on assumptions that must be evaluated empirically; therefore, backing evidence from theories, research, experience and original data, as well as from documents, regulations, and legal requirements (Bachman and Palmer 2010:102) is needed to support the assumptions. Inasmuch as the assumptions are tentative statements, data may not always yield supporting evidence, and certain factors may weaken the strength of the claim. These are presented as rebuttals that require additional evidence. Ultimately, the 'interpretive argument with the supporting evidence is the validity argument' (Chapelle 2012:19). Chapelle et al (Eds) (2008) put forth an exemplary argument for the Internet-based Test of English as a Foreign Language (TOEFL iBT), demonstrating how research can be defined through a systematic process of investigating specific inferences, and making clear 'how the validity argument can be questioned, weakened, limited or refuted by research that supports rebuttals' (Chapelle 2012:19).

Notwithstanding the strengths of the validity argument framework, establishing it as the sole approach to validation is not optimal. As Chapelle and Voss (2014:1,081) remarked, 'it is difficult to identify a one and only current view and practice in validation', and 'validity arguments themselves need to be evaluated for their completeness, coherence, and degree of support'. An area of concern with regard to completeness in this framework is the construct. Chapelle (2012) admits that while the argument-based validation process does not eliminate the need to define the construct, it downplays it – it is not the construct but the interpretive argument that serves as the basis of score interpretation. Xu (2018) reiterates the importance of the construct in validating automated scoring and feedback systems in particular. Importantly, Chapelle (2012:19) also explains that 'essential validation research is defined through a systematic process of examining the inferences in the interpretive argument rather than consulting a list of types of potential validity evidence'. For learning-oriented writing assessments like RWT, performance needs to be conceptualized and evaluated such that both the construct and specific sources of validity evidence take centre stage.

A complementary evidence-based approach that can enable such congruity is the socio-cognitive validity framework (Weir 2005).

The socio-cognitive validity framework

Positioned within the contemporary evidence-based paradigm for language assessment validation, Weir's (2005) socio-cognitive framework embodies traditional approaches that emphasize construct, content, and criterion types of validity as well as reliability, impact, and practicality. It also echoes Messick's (1989) thinking of accumulating validity evidence for proposed score-based claims and decision making. Distinct from the validity argument approach, which helps identify the weakest link in the argument and prioritize validation research, the socio-cognitive framework reshapes validity by treating it as an interactionalist representation where context, cognitive processing, and scoring interact with each other. Another distinguishing attribute is the more specific definition of the construct. Shaw and Weir (2007:2–3) describe a model for conceptualizing writing test performance, in particular, where the construct of writing is defined in terms of contextual and cognitive-based validity parameters, 'residing in the interactions between the underlying cognitive ability and the context of use'; it is 'not just the underlying traits of communicative language ability' but 'the result of the constructed triangle of trait, context and score (including its interpretation)'. Construct validity is then a conjunctive concept construed as a symbiotic relationship between cognitive, context, and scoring validity (Shaw and Weir 2007:7), encompassing the internal dimensions of assessment. The external dimensions are captured by criterion-related and consequential validity. Each type of validity is specified in terms of criterial individual parameters.

Cognitive validity refers to the representation of cognitive processes (e.g., macro-planning, micro-planning, monitoring, revising) involved in performing the writing task in a real-life context. These processes draw on the cognitive operations theorized in the writing scholarship grounded in cognitive psychology (e.g., Field 2004, Grabe and Kaplan 1996, Hayes and Flower 1980, Kellogg 1994). Context validity addresses real-life tasks in terms of the performance conditions of the target context. As Shaw and Weir (2007:63) put it, 'tests should approximate to "the performance conditions" of the authentic real-life context'. Its parameters include task and administration setting (e.g., purpose and uniformity of delivery, respectively), as well as linguistic demands (e.g., lexical resources, discourse mode, functional resources). Scoring validity, commonly known as reliability, is concerned with aspects of the assessment process such as scoring criteria, rater agreement, consistency in the application of rating scales, and stability over time. Scoring validity can be affected by inadequate performance conditions, and that can reduce construct-relevant variance.

Criterion-related validity assumes a relationship between test scores with other standards of performance that are believed to measure the same ability. Its parameters include cross-test comparability, comparison with different versions of the same test, and comparison with external standards. Consequential validity refers to external social consequences of test interpretation in terms of intended positive or negative effects. Parameters of relevance are washback on teaching and learning as well as broader influences on various stakeholders and educational practices.

In relation to RWT, the context and cognitive realizations of construct validity can be linked with the construct relevance of the rhetorical feedback generated for research article writing in various disciplines. Specifically, the evaluation of automated feedback that emphasizes textual and socio-disciplinary aspects of the context is related to the interaction between context validity (that concerns varying conventions across disciplines) and cognitive validity (that concerns the macro-planning and organization involved in the writing process). Similarly, the evaluation of automated feedback intended to facilitate monitoring, planning and revision of rhetorical traits falls in the area of cognitive validity.

Integrated validity frameworks

Although RWT was designed for use in different learning contexts, it targets a particular genre (RA) and a particular aspect of the writing construct (rhetorical). Hence, the interpretive argument articulates inferences associated with distinct propositional warrants within this scope. As shown in Figure 3, the argument consists of seven inferences that guide the type of validity research needed, each inference focusing on a specific facet of RWT feedback. Considering that argument-based approaches help 'establish a mechanism for combining different aspects of validity associated with automated scoring in an integrated fashion' (Williamson et al 2012:4), Figure 3 attempts to consolidate the interpretive and use argument by connecting it with the types of validity in the socio-cognitive framework. Note that the types of validity in the lower part of this figure should not be thought of as 'grounds' for the inferences they are linked to.

The first four inferences are linked with the internal context, scoring, and cognitive dimensions to show that they all contribute validity evidence about the construct the feedback is targeting. Context validity is associated with the domain definition inference because RWT's formative feedback must be, to the highest degree possible, representative of the target genre and the ability necessary for successful completion of the IMRD/C writing tasks in the target academic domains. Therefore, genre writing conventions should be included as descriptive parameters because they pose specific rhetorical and linguistic demands, which must be comprehensively described based

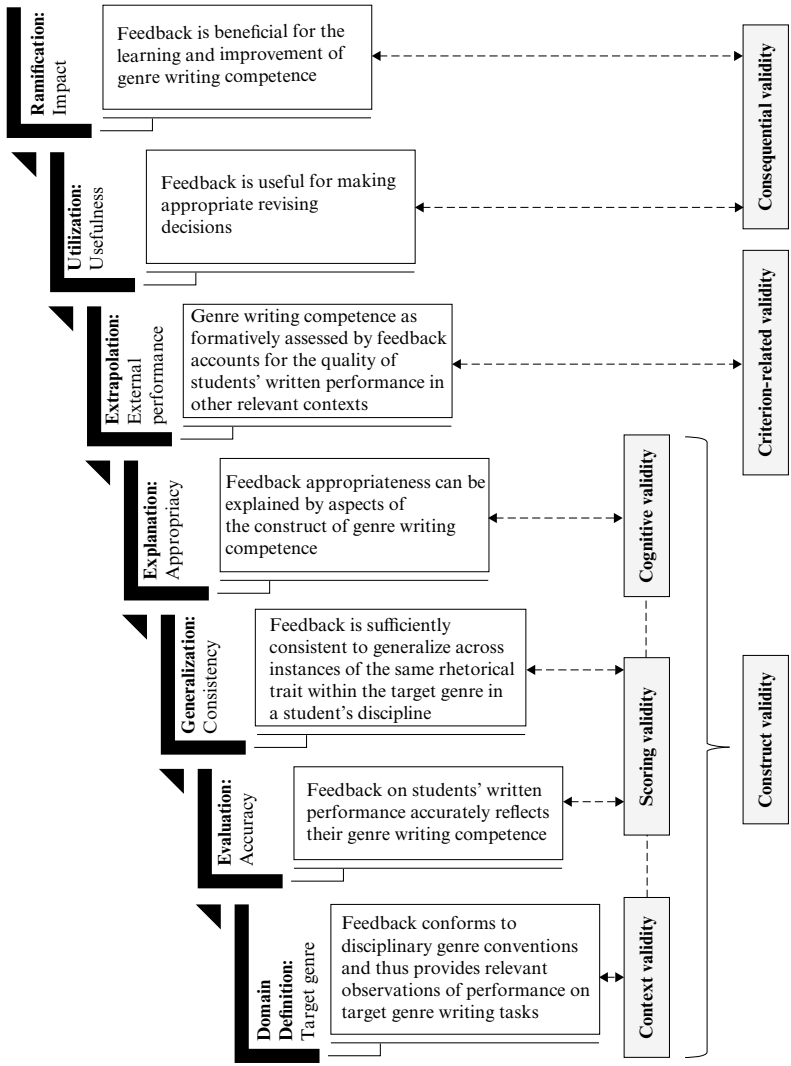


Figure 3 The interpretive and use argument for RWT

on authentic RA writing. Context validity, according to Weir (2005), is not separate but rather interconnects with scoring and cognitive validity.

Scoring validity arguably shares an interdependent relationship with the context/genre-based parameters. In testing, scoring criteria describe the level of writing performance that is required and are thus an essential part of the construct. For RWT, the criteria appropriate to the genre-writing tasks must conform to characteristic rhetorical conventions identified in authentic IMRD/C discourse. To convey appropriate information about the rhetorical quality of students' genre writing, the feedback must be, to the highest degree possible, accurate and consistent. Evidence for scoring validity can thus be gathered by investigating the warrants and assumptions articulated for the evaluation and generalization inferences that concern construct relevance and consistency of the feedback.

Cognitive validity in this symbiotic relationship specifies the cognitive processes that are activated by the task parameters (context validity) and that are necessary for the writing performance to be evaluated (scoring validity). The definition of cognitive validity takes into account cognitive writing models and theoretical explanations of skilled versus unskilled writing (Eysenck and Keane 2005, Scardamalia and Bereiter 1987), which is of direct relevance for RWT student users who may be strong writers but are novices to the RA genre. The feedback should stimulate cognitive and metacognitive operations needed to identify and resolve problems relating to both content and rhetoric, as do expert writers who approach the writing task bearing in mind their goals, audience, and genre features. To what extent RWT's feedback may do this appropriately can be learned through the lens of the explanation inference that accounts for feedback appropriateness in relation to the construct of the genre writing competence.

Criterion-related validity and consequential validity serve as evidence-based requirements external to the feedback. The former is relevant to the extrapolation inference, as it seeks to support predictions about the quality of performance in the real-world academic domain through evidence comparative to other measures of the same ability (e.g., advisor feedback, reviewer comments). Central to consequential validity in the case of RWT are two concerns – whether the feedback fosters successful revisions, and whether it has a positive impact on genre learning and writing improvement. These correspond to the utilization and ramification inferences, respectively, in the interpretive and use argument.

The space limits of this chapter preclude a full account of the warrants, assumptions, and backing evidence associated with each inference. Nonetheless, an example of unfolding the domain definition inference, which serves as the premise for the evaluation inference addressed below, may provide a sufficient degree of clarity about the logic behind the argument-based validation for RWT. Domain definition calls for a comprehensive

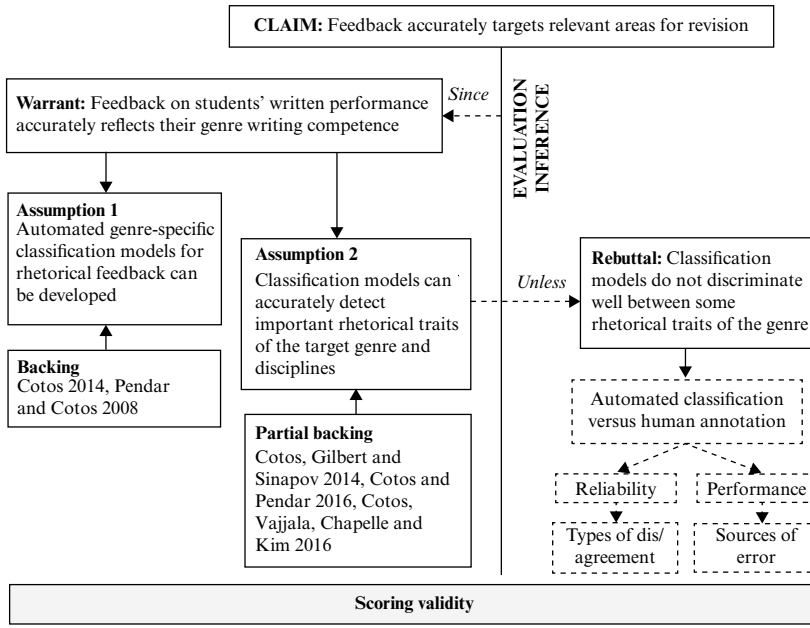
description of the RA genre, which represents the target domain where students' genre writing competence would be exercised. The claim that RWT supports IMRD/C writing tasks that are representative of RA writing in students' disciplines is accompanied by the warrant, assumption and backing elements of the argument, the general relationships among which were explained in the section 'Validity argument framework':

- *Warrant*: RWT feedback conforms to disciplinary RA genre conventions and thus provides relevant observations of performance from the students on target IMRD/C writing tasks.
  - *Assumption 1*: The rhetorical conventions of IMRD/C discourse can be identified and comprehensively described.
  - *Assumption 2*: The IMRD/C rhetorical conventions as used in authentic disciplinary discourse can be adequately represented.
- *Backing*: Large-scale corpus-based genre studies informed the design of RWT's feedback such that it epitomizes the IMRD/C rhetorical conventions and language use in 30 academic disciplines (Cotos, Huffman and Link 2015, 2017, Cotos, Link and Huffman 2016). This research describes specific rhetorical and linguistic demands in terms of moves and steps accomplished with specific functional language choices.

Investigating the evaluation inference

Once the domain definition inference was authorized with the above-mentioned warrant, it formed grounds for the evaluation inference detailed in Figure 4. The evaluation inference advances the argument with a warrant that RWT feedback on students' written performance accurately reflects their genre writing competence, thus being integral to scoring validity. This warrant rests on two assumptions. The first assumption concerns the feasibility of automated rhetorical analysis and is backed by research on the development of genre-specific rhetorical classification models (Cotos 2014, Pendar and Cotos 2008). The second assumption is about the quality of rhetorical analysis performed by the classification models and requires evidence obtained from empirical evaluations of the models. For the feedback to reflect students' genre writing competence, the classification models must accurately detect the move and step traits of IMRD/C discourse. A series of studies focused on the performance of the IMRD/C move and step classification models, yielding partial backing evidence (Cotos, Gilbert and Sinapov 2014, Cotos and Pendar 2016, Cotos, Vajjala, Chapelle and Kim 2016). On the one hand, algorithms and feature sets that were most stable and performed best for both move and step classification across disciplines and with discipline-specific corpora were identified. Specifically, word n-gram feature sets yielded better classification performance than

Figure 4 Evaluation inference in the interpretive argument for RWT



regular expressions and part-of-speech n-grams. As for learning algorithms, Support Vector Machines outperformed Naïve Bayes, Logistic Regression, and Random Forest models. (These machine learning algorithms for text classification rely on an inductive process that builds an automatic text classifier by learning the characteristics of the categories of interest from a set of pre-classified documents; see Sebastiani 2002). On the other hand, as detailed in the following section, the classification success for some rhetorical traits was not satisfactory.

Therefore, Figure 4 branches out a rebuttal that weakens the evaluation inference. The rebuttal motivated supplemental research to first explain why RWT classifiers may not discriminate well between some rhetorical traits of the genre and then to set an informed course of action. Henceforth, I only include data for the Introduction section to illustrate the types of evidence acquired through an exploratory study aimed to understand the problem at hand.

Grounds for rebuttal

The second assumption in Figure 4 received partial backing for the Introduction section because the performance measures of the classifier

varied; some rhetorical traits were discriminated well while others were not. In machine learning, the standard metrics of classification success are precision, recall, F1-score, and accuracy. Given that RWT feedback is generated based on classification output, such measures were obtained for each type of Introduction move and step trait numerated in Figure 2. Precision measured the proportion of sentences assigned to a trait that actually belonged to that trait, recall – the proportion of sentences belonging to a trait that were classified correctly; F1-score – the harmonic mean of precision and recall that measured the overall performance of the classifier for a given trait; and accuracy – the proportion of correctly classified sentences to the total number of classified sentences. Additionally, performance was measured through Cohen's kappa (Cohen 1960) that compared the classifier's accuracy against chance accuracy. This is recommended for training datasets that are not uniformly balanced, as was the case for the IMRD/C sub-corpora because some moves and steps of the genre are unavoidably sparse.

Through evaluations of RWT classification performance, it became clear that the problem of automated rhetorical analysis of moves and steps is more challenging compared to other text classification tasks (see Berger and Merkl 2005). Overall, the Cohen's kappa rates for step classification (.32) estimated by performing 10-fold cross-validation were lower compared to move classification (.47) (Cotos 2014). With regard to the classification of individual moves, Cotos and Pendar (2016) found that Move 2 (Identifying a niche), the sparsest yet the most argumentative move, is the most problematic. Move 2 had the lowest precision (59%), recall (37%), and F1-score (46%), tending to be misclassified as Move 1 (Establishing a territory). As for the step traits, 10 steps were classified well (e.g., Reviewing previous research, Outlining the structure of the paper) and seven were not (e.g., Raising general questions, Stating the value of present research). These results showed that RWT classifiers do not discriminate well between some rhetorical traits of the genre – a rebuttal that requires evidence-based explanations drawing on an understanding of how humans perform on this task.

Automated classification versus human annotation

In AWE validity research, evaluation has been anchored in analyses of reliability of the scoring engine, generally reporting correlations between human raters and computer scores (Shermis and Burstein (Eds) 2013). Multiple studies reported comparable human–computer and human–human agreements; measures such as weighted kappa ranged between .70 and .78 (Attali, Bridgeman and Trapani 2010), and Pearson's r was within the .70 and .91 range (Attali et al 2010, Rudner, Garcia and Welch 2006, Streeter, Bernstein, Foltz and DeLand 2011). Similar reliability indices were needed to address the rebuttal that emerged from RWT's classification performance.

It has been rightfully argued that establishing reliability by drawing on agreement-based calculations is necessary, but this should not be the sole source of evidence for automated scoring validity (Chung and Baker 2003). Apart from agreement with human scores, criteria associated with the evaluation inference for automated scoring can include standardized mean score difference, human scoring process and score quality, degradation from human–human agreement, threshold for human adjudication, and human intervention of automated scoring (Williamson et al 2012:5). The importance of various forms of human input necessary to meet these criteria cannot thus be underestimated. This is particularly true for RWT because it analyzes relatively abstract rhetorical concepts that can only be understood through human judgment. Moreover, human input is considered an explicit criterion of performance for automated scoring models and generally serves as the basis for their optimization (Williamson et al 2012), which is an implicit condition posed by the rebuttal. Therefore, classification output was examined vis-à-vis manual human annotation in order to accomplish two objectives:

1. To determine the degree of reliability between annotators, and between annotators and RWT classifiers, and understand the differences in agreement and disagreement.
2. To detect performance errors in human annotation and in RWT classification, and understand why they may occur.

Following these objectives, human and machine-generated types of data were analyzed to elucidate the nature of disagreement and misclassifications of the Introduction texts. Because RWT was designed to provide feedback on student writing but was trained on published texts, a small corpus of Introductions (30 texts) was compiled to include both expert texts published in peer-reviewed journals (15) and texts written by students (15). The student writers were native and non-native English language speakers (3 and 12, respectively). They were enrolled in a university-level research writing course and sought a graduate degree in Agronomy, Animal Science, Economics, Mechanical Engineering, or Sociology. All these disciplines are comprised in RWT. In this corpus, each discipline was represented by three published and three student texts. Because RWT's classifiers predict a move and a step for each sentence, all the texts were analyzed at sentence level. Table 1 presents the number of sentences in the corpus, which amounted to 940 in total.

Reliability and types of dis/agreement

In line with the first objective, all the texts in the corpus were analyzed using the output of RWT's Introduction classifier, the move and step classification data being extracted for each sentence. The same texts were also manually annotated in a similar manner by two experienced annotators, who were

Table 1 Number of sentences in the Introductions corpus*

Discipline	Published		Student	
	#	%	#	%
Agronomy	76	15	89	21
Animal Science	78	15	66	15
Economics	128	25	117	27
Mechanical Engineering	91	18	75	17
Sociology	132	26	88	20
Total	505	100	435	100

**Due to rounding not all percentages sum.*

trained research assistants previously involved in the prerequisite domain definition corpus-based research of the RA genre. They annotated all the texts independently, then discussed and resolved disagreement to produce a final, adjudicated annotation dataset. The inter-annotator reliability and annotator-classifier reliability were calculated using automated classifications and adjudicated annotations, the data being the same sentences. The numbers in Table 2 show simple percentage agreement, which provides the ratio of all the agreements over the total number of annotations, and Cohen's kappa, which accounts for expected chance agreement.

Two deductions applicable to both published and student texts can be drawn from these results:

- annotator–classifier reliability is lower than annotator–annotator reliability
- both annotator–annotator and annotator–classifier reliabilities are lower for step traits than for moves.

A general, almost default, explanation that humans understand meaning while the computer does not is certainly plausible, but hardly gratifying without knowing how exactly humans disagree when it comes to rhetorical annotation. Therefore, the next phase in this exploratory study focused

Table 2 RWT classification and human annotation reliability

Agreement	Trait	Percent agreement		Cohen's kappa	
		Published	Student	Published	Student
Adjudicated annotator–classifier	Move	68%	66%	.44	.42
	Step	42%	42%	.31	.33
Annotator–annotator	Move	94%	90%	.89	.84
	Step	80%	75%	.76	.71

on an inductive analysis aimed to identify possible levels of agreement and disagreement. This analysis was conducted at step level because the steps appeared to be more difficult to identify than moves for both the classifier and the annotators. To gain deeper insights, the annotators were instructed to annotate the same set of texts with multiple traits, as primary and secondary steps, if they thought a sentence had more than one step function. For example:

Sentence: 'Even as neighborhoods have integrated racially, neighborhoods and schools have become increasingly segregated by SES (CITATION).'

Annotation: primary M2_Step5 (Highlighting a problem) and secondary M1_Step3 (Reviewing previous research)

Identifying and annotating primary and secondary step functions captures the communicative intent more properly, as RA discourse is often multi-functional (Cotos et al 2015, Moreno and Swales 2018). More importantly, this procedure made it possible to discern specific themes from the two annotators' datasets and to better understand where and how their judgments diverged. Table 3 lists six types of agreement and disagreement along with definitions and examples.

The last two categories may not be as straightforward, so let us consider some examples. The first sentence on the next page exemplifies inverted agreement, as both annotators agreed on the steps, but their primary and secondary codes were switched. The second sentence exemplifies level

Table 3 Types of agreement and disagreement between annotators

Type	Definition	Example*	
		Annotator 1	Annotator 2
Agreement primary	The same primary step, regardless of the secondary step	P: M1_Step2	P: M1_Step2
Agreement secondary	The same secondary step, regardless of the primary step	S: M1_Step3	S: M1_Step3
Disagreement primary	Different primary step, regardless of the secondary step	P: M1_Step2	P: M1_Step3
Disagreement secondary	Different or additional secondary step, regardless of the primary step	S: M1_Step1 or no S	S: M1_Step3 or S: M2_Step8
Inverted agreement	The same steps, but different order of the primary and secondary steps	P: M1_Step2 S: M1_Step3	P: M1_Step3 S: M1_Step2
Level disagreement	The same step, but different primary or secondary level	P: M1_Step2	S: M1_Step2

* *Note: P and S stand for primary and secondary step functions, respectively.*

disagreement, as both annotators identified the same step (M1_Step3), but one annotator coded it as primary while the other annotator coded the same step as secondary.

Sentence: 'Traditional processes for DG production are based on continuous glycerolysis of fats and oils or direct esterification of glycerol with fatty acids with inorganic catalysts at higher temperatures and elevated pressure, which usually include high energy consumption, low yield, and poor product quality (CITATION).'

Annotator 1: M1_Step3 (Reviewing previous research – primary) and M1_Step2 (Providing general background – secondary)

Annotator 2: M1_Step2 (Providing general background – primary) and M1_Step3 (Reviewing previous research – secondary)

Type: Inverted agreement

Sentence: 'The increased incidence of both E. coli O157:H7 and STEC since the 1990s (24, 33) parallels the increased consumption of ground beef.'

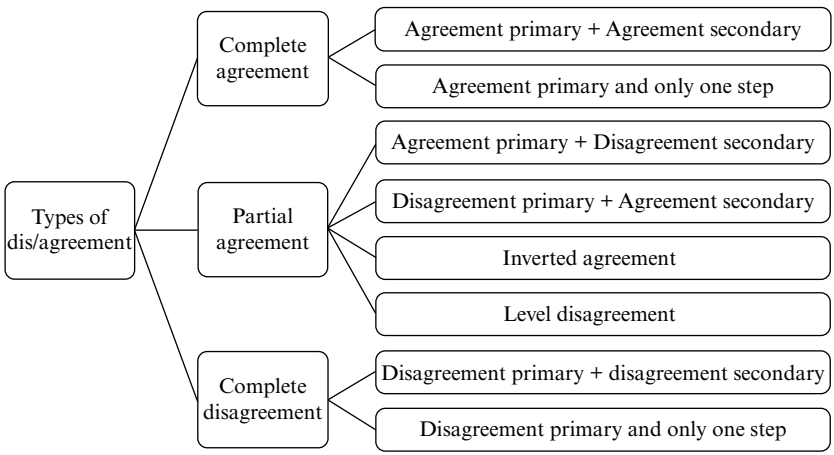
Annotator 1: M1_Step3 (Reviewing previous research – primary)

Annotator 2: M1_Step2 (Providing general background – primary) and M1_Step3 (Reviewing previous research – secondary)

Type: Level disagreement

More generally, the outcomes of the inductive analysis of agreement and disagreement can be shaped into a taxonomy, as in Figure 5. Now that these taxonomic categories have been identified and defined, they will be applied in

Figure 5 Taxonomy for analyzing agreement on rhetorical traits



subsequent research because they indicate that the concept of agreement on rhetorical annotation is not a dichotomic yes/no matter. Rather, estimates of human–human reliability should take into consideration a more nuanced annotation process because, by observing multiple step functions, it reflects how human readers may understand the nuances of communicative intent. This taxonomy will thus be used further to more precisely gauge the levels of annotator–classifier agreement on both move and step traits. Accounting for complete agreement as well as partial and level disagreement is expected to provide more adequate and acceptable reliability evidence.

Performance and sources of error

In addition to determining the differences in agreement and disagreement that can affect reliability, it is important to also gain an understanding of what may cause errors in automated rhetorical feedback. Therefore, following the second objective, human annotation and RWT classification performance were analyzed by computing confusion matrixes. A confusion (or error) matrix is a table containing correct and incorrect classification predictions summarized with frequency values for each class (see Kohavi and Provost 1998). The confusion matrixes were tabulated for steps for several reasons: a) inter-annotator reliability and annotator–classifier reliability were lower for steps than for moves; b) there are many more step classes than move classes, which makes the classification into steps a more difficult task; and c) errors are more apparent and more interpretable at step level.

The first confusion matrix, given in Figure 6, is for RWT classification compared to human-adjudicated annotation of sentences. The columns in this matrix represent the number of steps predicted by the classifier (classifier output), and the rows represent the steps assigned by the annotators (true class based on human input). The highlighted diagonal shows the number of correct classifications; the off-diagonal counts represent the number of classifications that were different from the adjudicated annotation. The matrix in Figure 7 summarizes the two annotators' performance in a similar manner, except that the columns and the rows represent the number of individual steps assigned by each annotator, and the diagonal line shows the number of sentences they annotated correctly, as in the adjudicated dataset.

Without going into much detail about the classifier and human performance on individual steps, the following general deductions can be drawn based on the confusion matrixes in Figures 6 and 7 (see all move and step names in Figure 2):

- both the classifier and the annotators can confuse a step with several other steps, e.g., M1_Step1 (Claiming centrality):
  - classifier – confused with six other steps: 2, 3, 4, 5, 8, 16 (Figure 6); annotators – confused with four other steps: 2, 3, 4, 5 (Figure 7)

Figure 6 Confusion matrix for RWT step classification

Annotators	Introduction step classifier																
	M1_Step1	M1_Step2	M1_Step3	M2_Step4	M2_Step5	M2_Step6	M2_Step7	M2_Step8	M3_Step9	M3_Step10	M3_Step11	M3_Step12	M3_Step13	M3_Step14	M3_Step15	M3_Step16	M3_Step17
M1_Step1	14	5	5	2	5	0	0	2	0	0	0	0	0	0	0	1	0
M1_Step2	12	87	16	3	11	0	5	4	3	1	0	0	0	3	2	2	0
M1_Step3	6	87	157	1	19	1	1	6	6	4	0	1	1	9	9	1	0
M2_Step4	4	7	4	12	1	1	1	1	1	1	0	0	0	0	1	1	0
M2_Step5	6	20	13	1	24	0	4	2	2	0	0	0	0	0	0	2	0
M2_Step6	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
M2_Step7	0	2	0	0	1	0	1	1	0	0	0	1	0	0	0	0	0
M2_Step8	0	8	2	1	0	0	4	6	1	0	0	1	0	0	0	4	0
M3_Step9	0	2	2	0	1	0	1	4	12	1	0	1	0	1	1	3	0
M3_Step10	0	3	3	0	0	0	0	0	2	11	0	0	0	3	1	2	0
M3_Step11	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
M3_Step12	0	4	3	0	2	0	2	0	4	0	0	2	0	0	2	0	0
M3_Step13	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	1	0
M3_Step14	0	32	19	0	4	0	1	3	26	0	0	0	0	18	3	8	0
M3_Step15	0	6	10	1	3	1	1	0	4	0	0	0	0	1	14	3	0
M3_Step16	2	3	2	1	0	0	0	4	3	0	0	0	0	0	0	7	1
M3_Step17	0	1	1	0	0	0	0	0	5	0	0	0	1	0	1	1	30

Figure 7 Confusion matrix for manual step annotation

Annotator 1	Annotator 2																
	M1_Step1	M1_Step2	M1_Step3	M2_Step4	M2_Step5	M2_Step6	M2_Step7	M2_Step8	M3_Step9	M3_Step10	M3_Step11	M3_Step12	M3_Step13	M3_Step14	M3_Step15	M3_Step16	M3_Step17
M1_Step1	21	4	4	1	2	0	0	0	0	0	0	0	0	0	0	0	0
M1_Step2	9	127	8	1	11	0	2	2	0	0	0	2	0	2	0	1	0
M1_Step3	2	30	271	2	13	0	0	1	0	0	0	0	0	3	1	0	0
M2_Step4	0	0	1	25	1	0	0	1	0	0	0	0	0	0	0	0	0
M2_Step5	1	2	5	9	41	1	0	0	0	0	0	0	0	1	1	0	0
M2_Step6	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
M2_Step7	0	3	1	0	0	0	0	0	0	0	0	3	0	0	0	0	0
M2_Step8	0	6	2	2	1	0	3	17	0	0	0	0	0	0	0	0	0
M3_Step9	0	1	0	0	1	0	0	0	16	5	0	3	0	14	2	2	3
M3_Step10	0	0	0	0	0	0	0	0	0	18	0	1	0	2	0	0	0
M3_Step11	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0
M3_Step12	0	0	0	0	0	0	0	0	0	0	0	9	0	2	0	0	0
M3_Step13	0	1	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0
M3_Step14	0	1	0	0	1	0	0	0	6	2	0	1	1	85	3	1	1
M3_Step15	0	0	0	0	0	0	0	0	0	0	0	0	0	4	40	1	0
M3_Step16	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0
M3_Step17	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36

- both the classifier and the annotators can confuse the same steps, e.g., M2_Step5 (Highlighting a problem):
  - classifier – 24 correct and 50 incorrect; both annotators – 41 correct, first annotator – 20 incorrect and second annotator – 30 incorrect
- both the classifier and the annotators can confuse steps pertaining to the same move (within boxes in Figures 6 and 7), e.g., Move 1 (Establishing a territory):
  - both annotators and the classifier confused M1_Step3 with M1_Step1 and M1_Step2
- both the classifier and the annotators can confuse steps pertaining to other moves (outside of boxes in Figures 6 and 7), e.g., M3_Step9 (Introducing present research descriptively):
  - classifier – confused with M1 steps (2, 3) and with M2 steps (4, 5, 8); annotators – confused with M1_Step1 and M2_Step5.

These patterns in erroneous performance by the classifier (which is trained to perform the classification task using n-gram features) and the humans (whose annotations may be confined by subjective interpretation of linguistic features despite clearly defined move/step categories and coding criteria) invite the supposition that there may be similarities in the possible sources of errors in human annotation and RWT classification. To verify this premise, qualitative analysis was conducted to acquire an understanding of possible sources of errors. The same dataset was analyzed inductively again, but this time with a focus on what may cause classification errors. For that, all misclassified sentences were extracted per step and manually analyzed.

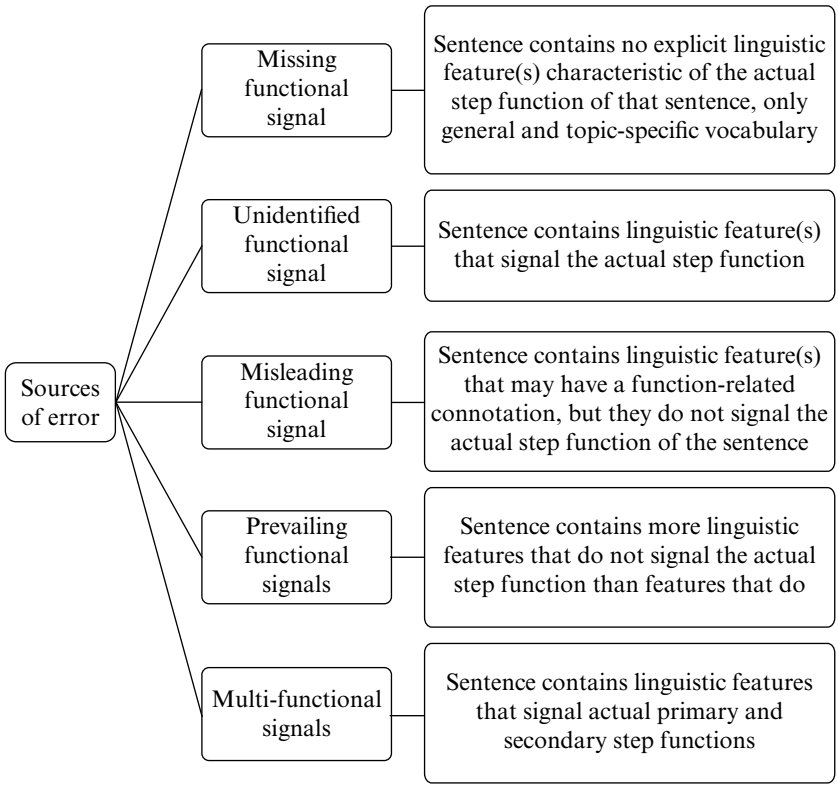
Both the annotators and RWT identify linguistic cues, or functional signals that are likely to be indicative of specific step functions (e.g., M2_Step4, a research gap conveyed by *have been rarely studied, no empirical data, it is currently unknown how*). In line with this shared aspect of human judgment and automated classification, functional language surfaced as the principal criterion in the analysis of misclassified sentences. The sentences in Table 4 exemplify classification errors, the italics showing the language choices that may cause confusion. While different sentences are provided as specific examples, it is not to say that the categories they exemplify are mutually exclusive.

The presence or absence of functional language cues such as those italicized above helped identify several themes that may explain the classifier's erroneous output in terms of potentially problematic types of functional signals. Figure 8 delineates them as possible sources of error, linking each category to tentative interpretations of why misclassifications may occur. If functional signals are missing, there may be no appropriate n-gram features in the training feature set, which is the only data used for classification (as

Table 4 Examples of functional signals in misclassified sentences

Sentence	Human annotation	Classification error	Possible source of error
'The potato trial and the three maize trials included sites and genotypes, whereas the two wheat trials included sites, genotypes, and years.'	M1_Step3 (Reviewing previous research)	M3_Step14 (Summarising methods)	Step-indicative functional language not present
'Diglyceride (DG), a naturally minor component of edible fats and oils, has attracted much attention over the past few years.'	M1_Step1 (Claiming centrality)	M1_Step2 (Providing general background)	*has attracted much attention over the past few years* not identified
'Nevertheless, the conical spouted bed reactor allows for minimizing the serious problem of defluidization caused by a very sticky reaction medium.'	M1_Step2 (Providing general background)	M2_Step5 (Highlighting a problem)	*serious problem* typical of M2_Step5
'Individuals immigrating to European countries, in particular non-Western immigrants, are found to be at higher risk of social exclusion, more often to be chronically poor, and to be over-represented in low-paid jobs CITATION.'	M1_Step3 (Reviewing previous research)	M2_Step5 (Highlighting a problem)	More functional language for M2_Step5 (*higher risk*, *poor*, *over-represented*, *low-paid*) than for M1_Step3 (*CITATION*)
'However, the crystal structure of the isolated nonphosphorylated ZAP-70 kinase domain complexed with staurosporine fails to show blockade of the catalytic site by the activation CITATION.'	M2_Step5 (Highlighting a problem) (primary) M1_Step3 (Reviewing previous research) (secondary)	M1_Step3 (Reviewing previous research)	Functional language for primary step (*However*, *fails*) and secondary step (*CITATION*)

Figure 8 Taxonomy for analyzing classification errors



opposed to annotation, as humans can derive functional meaning from the textual context when linguistic cues are absent). Furthermore, if functional signals indicative of the actual step function of a sentence are present but not identified, appropriate features may not have sufficient weight in classification. In the case of misleading and prevailing functional signals, inappropriate and dominant features may constrain classification. If a sentence contains signals of both primary and secondary functions, the problem is that the classifier predicts only one step. Functional signals that are missing, unidentified, misleading, prevailing, or multi-functional should be further analyzed to determine which of these types are more confusing for classification, and whether particular sources of error may be related to particular rhetorical traits. Therefore, the inductively identified error sources shown in Table 4 and taxonomically represented in Figure 8 will be used as heuristics in follow-up research that is needed to devise suitable approaches for improving the accuracy of rhetorical feedback.

Conclusion and implications

Designed as a genre writing enhancement tool, RWT offers feedback on the communicative moves and functional steps of the target RA genre, aligned with theoretical views on the rhetorical aspect of the writing construct and genre writing competence. As it is the case with other AWE and learning-oriented assessments, the approach to RWT validation draws on ideas from argument-based and socio-cognitive validation frameworks. This approach compels continued investigation in order to explain empirical findings that pose a rebuttal and to understand the confounding factors potentially affecting the accuracy of automated rhetorical feedback, the quality of which depends on the quality of classification performance.

Although limited to Introduction texts data, select evidence obtained in response to the rebuttal that RWT classifiers do not discriminate well between some rhetorical traits of the target RA genre provided a number of important insights derived from the exploration of automated classification versus human annotation. With regard to reliability, evidence suggests that the annotator–classifier reliability may be affected by a limitation in the classification approach (i.e., single-step classification), and that it may be lower than the annotator–annotator reliability due to a limitation in agreement coding (dichotomous agree or disagree). With regard to classification performance, the error analysis revealed that the classifier exhibited confusion/error patterns similar to human annotators, and that functional language (whether present or absent) may be key sources of misclassification of the rhetorical traits of texts. Considering these findings, it is recommended that a) the reliability measures for rhetorical analysis, both human and automated, should include different types of agreement and disagreement to account for the multi-functionality of research discourse, and b) classification performance measures should be accompanied by error analysis, as detecting the sources of errors can inform the optimization of classification models. The heuristic taxonomies devised through qualitative analysis (Figures 5 and 8) will be applied to pursue these two recommendations.

The taxonomy for analyzing human–classifier agreement is an important outcome, as measuring dis/agreement by type may ascertain how severe rhetorical misclassifications truly are and why the error rates seem to be higher for the classifier compared to humans. For example, the sentence 'Diglyceride (DG), a naturally minor component of edible fats and oils, has attracted much attention over the past few years' was annotated with M1_Step1 (Claiming centrality) as primary and M1_Step2 (Providing general background) as secondary steps, and classified as M1_Step2. This is an instance of less severe, partial disagreement because both the humans and RWT identified the same step, the difference being in that the annotators

marked it as secondary while RWT does not render secondary functions. This would in fact be possible because probability-based confidence levels are used to generate one feedback prompt for each sentence (e.g., probability of 90% and higher – *You are likely claiming centrality here*; probability of 75%–90% – *You may be claiming centrality*; probability of 50%–75% – *Are you claiming centrality*? and probability of 50% and lower – *Not sure what you are trying to do ... Can you be more explicit?*). Alternatively, the classification probability ranges could be regarded as distinguishing the computer-based primary and secondary steps for individual sentences, which could then be translated to multi-functional feedback. The system is not set up to do this, although it would be desirable for RWT to generate multi-step feedback in the future to more appropriately denote students' rhetorical ability. This would also be a more befitting representation of the construct because genre writing represents well-crafted argumentation where multiple step functions are intertwined at sentence level (Cotos et al 2015, Moreno and Swales 2018).

The taxonomy for analyzing classification errors is the second major outcome that can help recognize the types of functional signals possibly confusing the classifiers. Such evidence may explicate how specific linguistic features contained within a sentence contribute to its move/step classification, and also inform whether and how the feature sets and classification approach may need to be augmented for better construct representation in RWT feedback. An additional direction for classification experiments is leveraging what enhances human rhetorical annotation. As has been previously mentioned, the annotators comprehend functional meaning in light of the surrounding text while RWT's classifiers do not currently account for textual context. Complementary models (e.g., Fiacco, Cotos and Rose 2019) could be built to predict step sequences and determine the step functions of sentences considering the preceding and following moves and steps, especially in cases when functional cues are absent. Corpus-derived evidence from the domain definition research did indicate predictable step sequences, suggesting that some steps are more likely to follow certain steps, while some step transitions are highly unlikely (Cotos et al 2015).

Continued classification accuracy research is expected to lead to improved accuracy of RWT feedback, which may enhance the cognitive dimensions of construct validity and supply evidence to support the explanation inference in the interpretive argument. Similarly, improved feedback may enhance the external dimensions of consequential validity if it prompts students to make more appropriate revising decisions and helps them improve their genre writing. This, in turn, would serve as supporting evidence for the utilization and ramification inferences. In the meantime, RWT users need to be informed of existing strengths and weaknesses so that they are able to set rational expectations and exert autonomy when revising their writing with this tool.

By and large, the value of this work extends beyond the scope of RWT evaluation inference and scoring validity research, as it provides a more discerning picture of the complexity and difficulty of the rhetorical analysis problem. Additionally, it offers new heuristic taxonomies for rhetorically driven systematic inquiry of human and computer reliability and performance, thus contributing to the study of genre-based AWE. This field of practice, being still nascent, needs tools and procedures tailored to the analysis of human input and computer output in order to arrive at warrantable solutions.

References

Attali, Y, Bridgeman, B and Trapani, C (2010) Performance of a generic approach in automated essay scoring, *Journal of Technology, Learning, and Assessment* 10 (3), 4–16.

Bachman, L F and Palmer A S (2010) *Language Assessment in Practice: Developing Language Assessments and Justifying Their Use in the Real World*, Oxford: Oxford University Press.

Berger, H and Merkl, D (2005) A comparison of text-categorization methods applied to n-gram frequency statistics, in Webb, G I and Yu, X (Eds) *AI 2004: Advances in Artificial Intelligence. 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, December 4–6, 2004. Proceedings*, Lecture Notes in Computer Science 3339, Berlin: Springer, 998–1,003.

Chapelle, C A (2012) Validity argument for language assessment: The framework is simple…, *Language Testing* 29, 19–27.

Chapelle, C A and Voss, E (2014) Evaluation of language tests through validation research, in Kunnan, A (Ed) *The Companion to Language Assessment*, New York: Wiley, 1,081–1,097.

Chapelle, C A, Cotos, E and Lee, J (2015) Validity arguments for diagnostic assessment using automated writing evaluation, *Language Testing* 32 (3), 385–405.

Chapelle, C A, Enright, M K and Jamieson, J M (Eds) (2008) *Building a Validity Argument for the Test of English as a Foreign Language™*, New York: Routledge.

Chung, G K W K and Baker, E L (2003) Issues in the reliability and validity of automated scoring of constructed responses, in Shermis, M D and Burstein J C (Eds) *Automated Essay Scoring: A Cross-disciplinary Perspective*, Mahwah: Lawrence Erlbaum Associates, 23–40.

Cohen, J (1960) A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* 20 (1), 37–46.

Cotos, E (2014) *Genre-based Automated Writing Evaluation for L2 Research Writing: From Design to Evaluation and Enhancement*, London: Palgrave Macmillan.

Cotos, E (2015) AWE for writing pedagogy: From healthy tension to tangible prospects, *Writing and Pedagogy* 7 (2-3), 197–231.

Cotos, E (2016) Computer-assisted research writing in the disciplines, in Crossley, S A and McNamara, D S (Eds) *Adaptive Educational Technologies for Literacy Instruction*, New York: Routledge, 225–242.

Cotos, E and Pendar, N (2016) Discourse classification into rhetorical functions for AWE feedback, *CALICO Journal* 33 (1), 92–116.
Cotos, E, Gilbert, S and Sinapov, J (2014) NLP-based analysis of rhetorical functions for AWE feedback, in Colpaert, J, Aerts, A and Oberhofer, M (Eds) *Research Challenges in CALL*, *16th International CALL Research Conference Proceedings*, Antwerp: Linguapolis, Institute for Language and Communication, University of Antwerp, 117–123.
Cotos, E, Huffman, S and Link, S (2015) Furthering and applying move/step constructs: Technology-driven marshalling of Swalesian genre theory for EAP pedagogy, *Journal of English for Academic Purposes* 19, 52–72.
Cotos, E, Huffman, S and Link, S (2017) A move/step model for methods sections: Demonstrating rigour and credibility, *English for Specific Purposes* 46, 90–106.
Cotos, E, Link, S and Huffman, S (2016) Studying disciplinary corpora to teach the craft of discussion, *Writing and Pedagogy* 8 (1), 33–64.
Cotos, E, Vajjala, S, Chapelle, C A and Kim, H (2016) *Computational analysis of methods sections in a corpus of scientific articles*, poster presented at the American Association of Corpus Linguistics and Technology for Second Language Learning Conference, Ames, IA.
Deane, P (2013) On the relation between automated essay scoring and modern views of the writing construct, *Assessing Writing* 18, 7–24.
Eysenck, M and Keane, M (2005) *Cognitive Psychology* (Fifth edition), Hove: Psychology Press.
Fiacco, J, Cotos, E and Rose, C (2019) *Towards enabling feedback on rhetorical structure with neural sequence models*, paper presented at the 9th International Conference on Learning Analytics & Knowledge, Tempe, Arizona, March 2019.
Field, J (2004) *Psycholinguistics: The Key Concepts*, London: Routledge.
Gleason, J (2013) An interpretive argument for blended Spanish tasks, *Foreign Language Annals* 46 (4), 588–609.
Grabe, W and Kaplan, R B (1996) *Theory and Practice of Writing: An Applied Linguistics Perspective*, London: Longman.
Gruba, P, Cárdenas-Claros, M S, Suvorov, R and Rick, K (2016) *Blended Language Program Evaluation*, London: Palgrave Macmillan.
Hayes, J R and Flower, L S (1980) Identifying the organisation of writing processes, in Gregg, L W and Steinberg, E R (Eds) *Cognitive Processes in Writing*, Mahwah: Lawrence Erlbaum Associates, 3–30.
Jones, N and Saville, N (2016) *Learning Oriented Assessment*: *A Systemic Approach*, Studies in Language Testing Volume 45, Cambridge: UCLES/ Cambridge University Press.
Kane, M T (2011) Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010, *Language Testing* 29 (1), 3–17.
Kane, M T (2013) Validating the interpretations and uses of test scores, *Journal of Educational Measurement* 50 (1), 1–73.
Kellogg, R T (1994) *The Psychology of Writing*, New York: Oxford University Press.
Kohavi, R and Provost, F (1998) Glossary of terms, *Machine Learning* 30, 271–274.
Li, Z, Link, S, Ma, H, Yang, H and Hegelheimer, V (2014) The role of automated writing evaluation holistic scores in the ESL classroom, *System* 44 (1), 66–78.

Link, S and Li, J (Eds) (2018) *Assessment Across Online Language Education*, CALICO Monograph Series, CALICO Monograph Series Volume 16, San Marcos: CALICO.

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

Moreno, A I and Swales, J M (2018) Strengthening move analysis methodology towards bridging the function-form gap, *English for Specific Purposes* 50, 40–63.

Pendar, N and Cotos, E (2008) *Automatic identification of discourse moves in scientific article introductions*, paper presented at the third ACL workshop on the innovative use of NLP for building educational applications, Columbus, Ohio, June 2008.

Perelman, L (2012) Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES), in Bazerman, C, Dean, C, Early, J, Lunsford, K, Null, S, Rogers, P and Stansell, A (Eds) *International Advances in Writing Research: Cultures, Places, Measures*, Fort Collins: WAC Clearinghouse/Anderson, 121–131.

Purpura, J C and Turner, E (2014) *Learning-oriented Assessment in Language Classrooms: Using Assessment to Gauge and Promote Language Learning*, New Perspectives on Language Assessment Series, London: Routledge.

Ranalli, J, Link, S and Chukharev-Hudilainen, E (2017) Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation, *Educational Psychology* 37 (1), 8–25.

Rudner, L M, Garcia, V and Welch, C (2006) An evaluation of the IntelliMetric Essay Scoring System, *The Journal of Technology, Learning, and Assessment* 4 (4), 3–18.

Scardamalia, M and Bereiter, C (1987) Knowledge telling and knowledge transforming in written composition, in Rosenberg, S (Ed) *Advances in Applied Psycholinguistics Volume 2: Reading, Writing and Language Learning*, Cambridge: Cambridge University Press, 142–175.

Sebastiani, F (2002) Machine learning in automated text categorization, *ACM Computing Surveys* 34, 1–47.

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

Shermis, M D and Burstein, J C (Eds) (2013) *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, New York: Routledge.

Streeter, L, Bernstein, J, Foltz, P and DeLand, D (2011) *Pearson's Automated Scoring of Writing, Speaking, and Mathematics. White Paper*, available online: images.pearsonassessments.com/images/tmrs/PearsonsAutomatedScoringofWritingSpeakingandMathematics.pdf

Swales, J M (1981) *Aspects of Articles Introductions*, Aston ESP Reports No. 1, Birmingham: Aston University.

Swales, J M (1990) *Genre Analysis: English in Academic and Research Settings*, Cambridge: Cambridge University Press.

Tardy, C M (2009) *Building Genre Knowledge*, West Lafayette: Parlor Press.

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

Williamson, D M, Xi, X and Breyer, F J (2012) A framework for evaluation and use of automated scoring, *Educational Measurement: Issues and Practice* 31 (1), 2–13.
Xu, J (2018) Measuring "Spoken Collocational Competence" in Communicative Speaking Assessment, *Language Assessment Quarterly* 15, 255–272.
Yang, H and Cotos, E (2018) Innovative implementation of a web-based rating system for individualizing online English-speaking instruction, in Link, S and Li, J (Eds) *Assessment Across Online Language Education*, CALICO Monograph Series Volume 16, San Marcos: CALICO, 167–183.

13 Building an auto-marker for assessing spontaneous L2 English speech

Kate Knill
University of Cambridge, UK

Mark Gales
University of Cambridge, UK

Introduction

With the growth in demand globally for learning English as an additional (L2) language, there is considerable interest in methods to automatically assess a learner's spoken language proficiency, both for qualifications and to provide immediate feedback to a learner's speaking practice. A lot of time and effort has been expended into developing the construct for human-marked tests which we do not want to throw away. Thus the aim of a spoken language assessment (SLA) auto-marker is to predict the score that a human examiner would assign to oral test responses under the same construct.

Initial SLA auto-marking systems tended to focus on assessing pronunciation and fluency using a *selected response* approach (e.g., Cucchiarini, Strik and Boves 1997, Franco et al 2000). Here, the learner is generally guided to read a given sentence and their pronunciation is scored against a metric such as goodness of pronunciation (Witt 1999). This prevents a learner from demonstrating their full spoken language ability (Khabbazbashi, Xu and Galaczi 2021) but is still the dominant mode of automatic SLA. To assess a wider range of linguistic features, elicitation of longer, spontaneous responses is needed, i.e., a *constructed response* approach. This is typically achieved in human-examined tests through monologic *open speaking* tasks where a candidate speaks for 10–60 seconds in response to a prompt. Automation of the constructed response testing process is more challenging than for selected response tests. Spontaneous speech contains disfluencies (hesitations, false starts, etc.) and is rarely fully grammatical or in clean sentences. Combined with the accented speech produced by learners, interpreting what a learner has said can be difficult even for human examiners. This is compounded when the speech signal is degraded due to background noises, such as a noisy classroom, and/or

distortions caused by the recording channel. This chapter describes an auto-marker for monologic[1] free speaking prompt-response tests that handles the above challenges to match human examiner performance.

Related work

In constructed spoken response testing we do not know what the candidate has said so automatic speech recognition (ASR) is required to transcribe the audio. To simplify the ASR task, automatic constructed response assessment was applied initially to short answer tasks requiring factual information (e.g., Leacock and Chodorow 2003) and tasks where the speech elicited was highly predictable (e.g., Bernstein 1999). The first system to tackle fully free speaking proficiency assessment was the Educational Testing Service (ETS) SpeechRater™ (Higgins, Xi, Zechner and Williamson 2011, Xi, Higgins, Zechner and Williamson 2008, Zechner, Higgins, Xi and Williamson 2009). For example, it has been applied to the high-stakes Internet-based Test of English as a Foreign Language (TOEFL iBT) which contains candidates' responses to both textual and audio-visual stimuli. The grader models use features based on audio and fluency as in earlier pronunciation-focused assessment (Cucchiarini et al 1997, Franco et al 2000), and features related to pronunciation, grammatical accuracy and ASR confidence. The accuracy of the ASR systems used to generate features has been improved significantly in recent years following the introduction of deep neural networks (DNN)[2] (Hinton et al 2012), for example (Cheng, Chen and Metallinou 2015, Hu, Qian, Soong and Wang 2015, Knill et al 2018, Metallinou and Cheng 2014, Tao, Ghaffarzadegan, Chen and Zechner 2015, van Dalen, Knill and Gales 2015).

A variety of machine learning classifier and regression models have been proposed for SLA graders including: regression models (Evanini and Wang 2013), support vector machines (SVMs) (Loukina, Zechner, Chen and Heilman 2015), and Gaussian Processes (van Dalen et al 2015). van Dalen introduced the concept of predicting the level of uncertainty of the grader in its predicted score to enable deployed systems to decide whether the automatic grade can be returned 'as is' to the candidate or if additional marking by a human is required. A problem with these models is that they need more memory and computation as the amount of training data increases, leading to impractical systems. This is not an issue with neural networks so a number

1 L2 English tests may also have dialogic, conversational, components. The additional challenges of automatically assessing dialogic speaking tests mean that these are outside the scope of this chapter.

2 The creation of foundation models for ASR, such as wav2vec2.0 and OpenAI's Whisper, has yielded further significant improvements. These models are out of the scope of this chapter.

of DNN-based grader approaches have been proposed (Hu, Richmond, Yamagishi and Latorre 2013, Malinin, Ragni, Gales and Knill 2017, Qian et al 2019).

Open speaking auto-marking

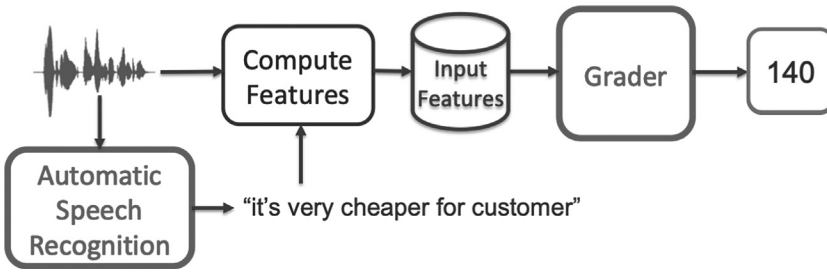
In open speaking tests (e.g., Linguaskill, and parts of IELTS and TOEFL) the candidate is asked to respond to a series of prompts over a range of tasks such as talking about their home town or contrasting two pictures. Candidates are expected to talk for between 10 and 60 seconds depending on the task. This allows them to demonstrate their ability to use their spoken language skills effectively. The learner's proficiency in coherence and discourse management, language resource and achievement of the communicative goal are assessed, in addition to their pronunciation and fluency.

Human examiners follow a standardised marking scheme with marks awarded based on task achievement; coherence/discourse management; language resource; pronunciation; hesitation/extent. Descriptions of what is expected of a candidate are provided for each level, e.g., 'there is an adequate range of grammar and vocabulary which is sufficiently accurate to deal with the tasks'. Examples of speech at different levels are provided to the examiners to help them differentiate between the grades during examiner training and standardisation. It would be very difficult to write down rules for all the aspects of assessment that examiners have to take into account. Instead, auto-marking systems are based on machine learning (ML), where the rules behind examiner judgements are inferred and learned from data.

ML systems infer (predict) one or more labels from a set of input data. For example, an auto-marking system predicts the grade or score of a speaker given their audio, and possibly metadata such as their L1. The core of an ML system is a statistical model whose model parameters are trained on examples of data from the same or a similar task as the intended prediction task. In a supervised training approach, the ML model predicts the labels for the training data where the target labels are known. Over a number of iterations, a training algorithm guides the ML model parameters towards minimising the error between the model's predictions and the targets. As it is a statistical model, the more data that is seen in training, the more robust the model parameter estimates. The wider the variety of training data, such as data from a range of L1 and/or proficiency levels, the better the model will generalise to unseen candidate data.

Figure 1 shows the basic components of an open speaking auto-marker. One of the challenges of assessing open speech is that what the candidate said is unknown. In a feature-based auto-marker simply extracting information from the audio signal, without any knowledge of the word sequence, does

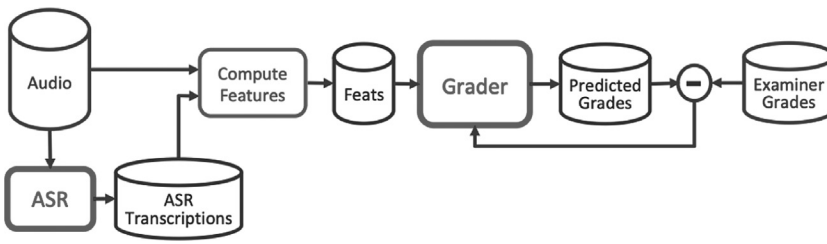
Figure 1 Open speaking auto-marker



not yield enough information for the grader to predict the grade accurately.[3] The first stage in assessing open speech is, therefore, to run an ASR system to transcribe the input waveform. The output from the ASR is used to 'structure' the audio to enable rich input features to be extracted from both the recognised text and the audio. In addition, detailed information can be extracted from this structured audio to provide feedback to the speaker, for example the pronunciation and grammatical correctness of their speech.

The auto-marking structure introduced in Figure 1 is a highly flexible framework but the performance of the system is dependent on the accuracy of the speech recognition system. Thus, approaches for open speaking assessment need to *minimise* the errors in the speech recognition, and *mitigate* the impact these errors have on the auto-marking and feedback. Figure 2 illustrates one form of error mitigation that is used: the automatic grader is trained based on transcriptions from the ASR system rather than manually transcribed texts. This means that the auto-marker is trained to predict the candidate's grade given a potentially erroneous transcription, capturing the relationship between grade and consistent errors that occur in the ASR output, for example the consistent confusion of two words. It is important

Figure 2 Framework for training the grader component



3 Foundation model-based neural speech auto-markers have been shown to be able to mark directly from the audio signal (Bannò and Matassoni 2023, Bannò, Knill, Matassoni, Raina and Gales 2023).

that the training data used to train the ASR system and that used to train the grader system are kept distinct. This ensures that the grader training mode exactly mimics the prediction mode.

The next sections describe the ASR and grading components in more detail, followed by evaluation of the auto-marking system.

Non-native speech recognition

ASR systems have become commonplace over the last decade, particularly with the rise in virtual assistants for smart phones and home devices such as Apple's Siri, Amazon's Alexa, Microsoft's Cortana and Google's Assistant. The performance of ASR has improved rapidly to support this expansion. This progress has been driven by both advances in the underlying technology and the availability of large quantities of data to train the systems. Despite these advances, recognition of non-native speech still remains a significant challenge, as anyone who has tried to use an assistant in another language will testify. This is especially true for multi-level spoken language assessment where the system is required to recognise speech from a wide range of L1s and over a broad range of language proficiency. In addition, recording conditions are often noisy due to, for example, background speech from other candidates taking the test at the same time, and background noise such as air conditioners or street sounds. To optimise performance, an ASR system customised for learner non-native speech recognition is required.[4]

Automatic speech recognition

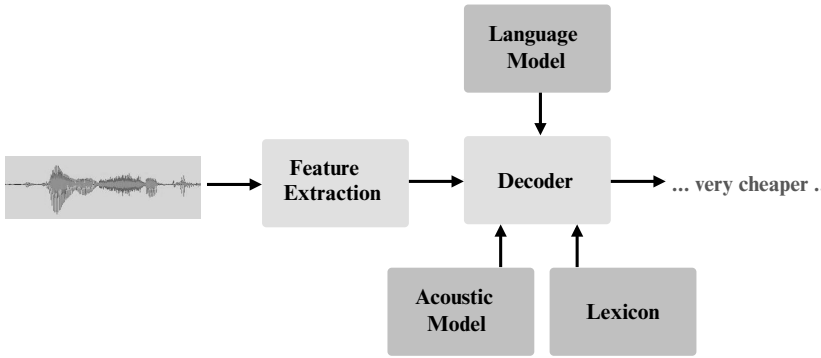
The aim of ASR is to convert a speech waveform into text; the word sequence associated with the waveform. Figure 3 shows the standard structure for an ASR system.

The first step[5] in this process is *feature extraction*. Speech is typically recorded at a 16kHz sampling rate, i.e., there are 16,000 samples per second, one every 0.0625 milliseconds (ms). Operating on each sample would require too much computation so a compact set of features is derived for the speech samples every 10ms, $\boldsymbol{x}_{1:T} = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$, where T is the number of input frames. The set of input features is chosen with the aim of extracting pertinent information from the speech signal whilst ignoring anything that does not

4 ASR foundation models, such as OpenAI's Whisper and Meta's MMS, offer considerably better out-of-the-box performance for L2 English, and are improving all the time. They still require some adjustment for use in SLA (Ma, Qian, Gales and Knill 2023).

5 In recent years there has been increasing interest in 'end-to-end' systems where the various stages, and modules, in Figure 3 are merged into one stage which is trained. Though this has some very nice properties, it requires large amounts of data to be available to train the system, which is not possible for current non-native ASR systems.

Figure 3 Elements of a standard automatic speech recognition system



contribute to the recognition result. The number of features per frame is typically in the range of 39 to 80. Each feature vector is generally computed on 25.6ms of speech, which roughly corresponds to periods for which human speech is stationary. Overlapping the frames by sampling every 10ms is done to handle the fact that we don't know when stationary periods start. The extracted features are fed to a decoder which generates the text sequence, $w_{1:L} = w_1, \ldots, w_L$.

Three distinct sources of information are used by the ASR system to decode the underlying text sequence, all of which are impacted by the waveform being generated by a non-native speaker. The sources of information are:

- *Acoustic Model*: this component models each of the acoustic realisations of the sounds of English. As it would be impossible to model every word in the English language,[6] the acoustic models typically represent sub-word units such as phones. The decoder generates the likelihood of the acoustic vectors from the feature extraction being generated by a word sequence, $p(\boldsymbol{x}_{1:T}|\boldsymbol{w}_{1:L})$. The lexicon is used to map the word sequence to the appropriate sequence of sub-word acoustic models.
- *Language Model*: many sounds in English are acoustically confusable, e.g., /m/ and /n/, so a language model is used to help the ASR decoder distinguish between similar sounding word sequences e.g., 'recognise speech' vs 'wreck a nice beach'. The language model generates the probability of any word sequence, $P(\boldsymbol{w}_{1:L})$.
- *Lexicon*: the lexicon component is required to map from a word in the word sequence to the sub-word-units that are used in the acoustic

6 Not least because new words are being invented all the time.

model. Thus it performs the following mapping $w_i \rightarrow s_1^{(i)}, \ldots, s_K^{(i)}$, where each sub-word unit $s_{1->k}^{(i)}$ has an associated acoustic model. Two forms of sub-word unit will be discussed in this chapter.

phones	CHEAPER	→	/ch/ /iy/ /p/ /ax/
graphemes	CHEAPER	→	/c/ /h/ /e/ /a/ /p/ /e/ /r/

As we cannot move our jaws and lips etc. instantaneously when we speak the sounds are affected by their context and co-articulation occurs. The context of the sub-word unit, such as the sounds before and after, triphones or tri-graphemes, and context markers such as the position of the sub-word unit within the word, may be used to boost the discrimination of the acoustic models.

The most common form of decoder selects the word sequence $\hat{\boldsymbol{w}}$ that has the maximum likelihood of generating the observation sequence, $\boldsymbol{x}_{1:T}$, i.e., what is the most likely series of words that the speaker has said. Mathematically this can be expressed as

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} \{P(\boldsymbol{w}|\boldsymbol{x}_{1:T})\} = \arg\max_{\boldsymbol{w}} \{P(\boldsymbol{w})p(\boldsymbol{x}_{1:T}|\boldsymbol{w})\} \qquad (1.1)$$

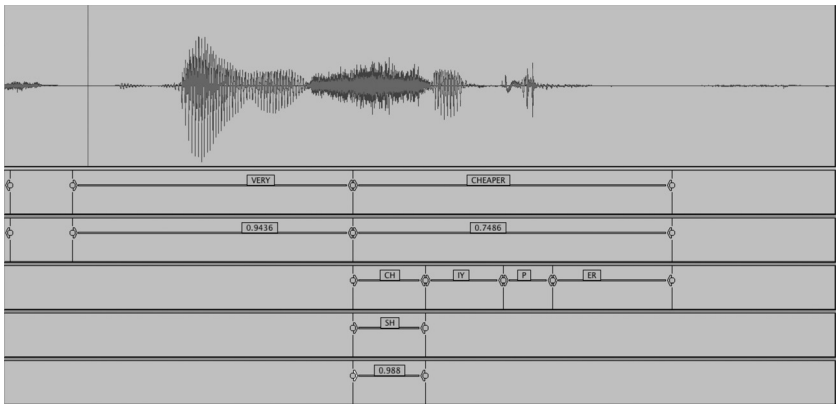
The framework described above, sometimes referred to as a generative classifier, has been extensively used for speech recognition for English and a wide range of other languages. Applying the above process to non-native speech recognition poses a number of challenges. To illustrate some of the issues, take an example sentence from a French L1 speaker.

Words	IT'S		VERY				CHEAPER		FOR	CUSTOMER	
Phones	…	/v/	/eh/	/r/	/iy/	/sh/	/iy/	/p/	/ax/	…	…

This example illustrates two challenges for ASR systems. First, the pronunciation of the word CHEAPER is not standard, so a common substitution for French L1 speakers of the phone /ch/ with /sh/ has occurred. This impacts the lexicon. Second, the sentence is not grammatically correct. A native speaker of English would more likely say IT'S MUCH CHEAPER FOR CUSTOMERS. This poses a problem for the language model as the spoken sentence is unlikely to occur in the standard language model training data. In addition, all the phones carry the French accent to a certain extent and the learner may use prosodic phrasing from French, where stress is placed on the final syllable of a word whereas the first syllable would be stressed in CHEAPER and CUSTOMER by a native speaker. All these factors need to be captured within the ASR components.

Figure 4 shows the sort of information that can be derived from the ASR system. A screenshot from an *audacity* project is shown. The first block shows the waveform that is fed into the ASR system. The second block shows part of the word sequence, and associated start and end times for each word,

Figure 4 Example output that can be obtained from an ASR system



that the ASR system generates. In addition to the word sequence, the third block shows the confidence score, which is how confident the ASR system is that the word has been recognised correctly. The fourth block shows the standard phonetic sub-word sequence from the lexicon associated with the word `CHEAPER`. The fourth and fifth blocks show the sort of pronunciation feedback that can be derived, showing the pronunciation error /sh/ and the probability of this error.

Acoustic and language models

Speech recognition systems are usually trained in a *supervised* training fashion, that is the training data consists of a set of paired training examples where each input sample is labelled with a desired output. Here the training data comprises the acoustic features derived from a corpus of speech (input) and the transcriptions of what was said, the word sequence (output labels). Thus the training data, D, comprises pairs of data: the observation sequence, $\boldsymbol{x}_{1:T}$, and the word sequence, $\boldsymbol{w}_{1:L}$. This data is used to train the acoustic model and the language model. The lexicon is assumed known.

The standard form of language model used in speech recognition is the n-gram language model. This counts the frequency of sequences of n words in training texts. The more common the sequence, the higher the probability, e.g., 'dog bites man' is much more likely than 'man bites dog'. Here the probability of the L length word sequence $\boldsymbol{w}_{1:L}$ is approximated, taking the example of a two-gram language model:

$$P(\boldsymbol{w}_{1:L}) = P(w_1)\prod_{i=2}^{L} P(w_i|\boldsymbol{w}_{1:i-1}) \approx P(w_1)\prod_{i=2}^{L} P(w_i|w_{i-1}) \tag{1.2}$$

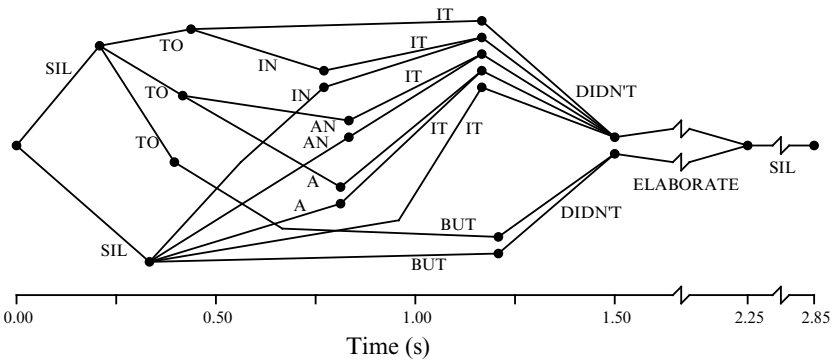
Most ASR systems use a trigram language model for decoding. For example, for the phrase 'The cat sat ...', the trigram probabilities for some words following 'cat sat' P (w_i | cat sat) might look like:

$$\begin{aligned} \text{P(on | cat sat)} &= 0.50 \\ \text{P(to | cat sat)} &= 0.01 \\ \text{P(rabbit | cat sat)} &= 0.000001 \end{aligned}$$

Although simple, the n-gram language model has proven to be highly effective in modelling spoken language. N-grams outperform more complex models based on models of grammar as we do not tend to speak precisely. They are also very computationally efficient to apply. Since the n-gram model focuses on a short word history, however, it is unable to model longer span word choices. Additional accuracy has been achieved by applying deep learning to train recurrent neural network language models (RNNLM) (Mikolov, Karafiát, Burget, Černocký and Khudanpur 2010). The best performing RNNLMs incorporate both a longer history and look ahead at succeeding words (Chen, Liu, Ragni, Wang and Gales 2017, Xiong et al 2017). They cannot be applied directly to decoding so an n-gram decode is typically run and alternative hypotheses for what was said stored in a lattice format (Figure 5). The RNNLM is then applied to rescore these hypotheses, replacing the n-gram LM scores with those from the RNNLM. The new best path through the lattice is used as the ASR output. RNNLMs require more data for training than a n-gram so are not always feasible.

To train the language model the word sequence, $\boldsymbol{w}_{1:L}$, is used. However, the quantity of data normally available from the audio training data is often very limited. Typically, native speakers speak at a rate of about 10,000 words an hour. Thus, for a training corpus of 1,000 hours of training data, this would yield approximately 10 million words. The typical ASR vocabulary (number

Figure 5 Example of ASR hypothesis lattice



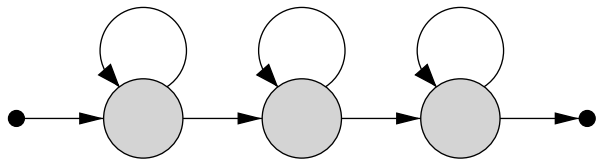
of entries in the lexicon) is about 65,000. This means that the amount of data to train each of the language model entries $P(w_i|w_{i-1})$ in the example above is often limited, or non-existent. To address this problem, language models are often trained on large quantities of text data collected from scraping the web, for example. Although this increases the quantity of data available, it is not necessarily closely matched to the word sequences produced in natural speech. As illustrated in the previous section, this is expected to be an even greater concern for non-native learners of English who will not always use grammatically correct sentences.

In addition to the language model, an acoustic model needs to be trained. This makes use of both the audio and transcriptions. The most common form of acoustic model used is based on Hidden Markov Models (HMMs). There is a separate HMM for each sub-word unit. A standard topology for the HMM is strict left-to-right; an example for a three-state model is shown in Figure 6. To compute the likelihood of a particular sequence of acoustic features, a sequence of HMMs, as defined by the lexicon, are connected together to form a model of a word sequence. The probability of an observation sequence is then given by:[7]

$$p(\boldsymbol{x}_{1:T}|\boldsymbol{w}_{1:L}) = \sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_{\boldsymbol{w}_{1:L}}} p(\boldsymbol{x}_{1:T}|\boldsymbol{\theta}_{1:T})\,P(\boldsymbol{\theta}_{1:T}) = \sum_{\boldsymbol{\theta}_{1:T}\in\boldsymbol{\Theta}_{\boldsymbol{w}_{1:L}}} \prod_{t=1}^{T} P(\theta_t|\theta_{t-1})p(\boldsymbol{x}_{1:t}|\theta_t) \quad (1.3)$$

where $\boldsymbol{\theta}_{1:T}$ is a T length state sequence and $\boldsymbol{\Theta}_{\boldsymbol{w}_{1:L}}$ is the set of all valid state sequences defined by using the lexicon to map the word sequence to a sub-word sequence of HMMs. There are a range of possible acoustic models that have been used to predict the HHM state probabilities, $p(\boldsymbol{x}_{1:t}|\theta_t)$, of which deep learning models dominate (Gales and Young 2007, Hinton et al 2012, Yu and Li 2017). The quantity of training data determines the number of sub-word units, and associated states, that are trained for an acoustic model. For details of the decisions made for balancing the number of these units with data see Gales and Young (2007).

Figure 6 Typical sub-word unit topology



7 Again end-to-end systems are not being described, and bi-directional models are not being considered.

Lexicon

In a good-quality phonetic lexicon such as Combilex (Richmond, Clark and Fitt 2009) and CMUdict (*The CMU Pronouncing Dictionary*, no date), many of the pronunciations will have been handcrafted by phoneticians to take into account the vagaries of English. They are typically produced to cover a single accent or variant, e.g., Received Pronunciation (RP) English or American English, i.e., not non-native pronunciations. If a word is not found in a lexicon, then we must automatically produce a pronunciation for the word using a grapheme-to-phoneme (G2P) conversion tool such as Sequitur (Bisani and Ney 2008) and Phonetisaurus (Novak, Minematsu and Hirose 2015). Since English does not follow a simple mapping from orthography to phones (e.g., bough, cough, dough) G2P-derived pronunciations tend to be less accurate. This is particularly true for proper nouns, including place and company names, which are most likely to converge from standard 'rules'. These sorts of terms are common in open speaking tests where candidates are often asked to talk about something relating to themselves, yielding a vast vocabulary in the training data across responses from around the world.

By contrast, graphemic lexicons for English are simple to craft – each of the 26 letters of the alphabet is a sub-word unit. The graphemic pronunciation is therefore the letters of the word. Apostrophes can be marked on the preceding letter or first letter of the word if at start of the word, which can be used in acoustic model building. We do lose, however, some richness in modelling moving from the over 40 English phones to the 26 graphemes.

To provide further refinement in a lexicon and more fine-grained acoustic models, the position of a sub-word unit in a word can be marked (beginning B, middle M, end F) (Gales, Knill and Ragni 2015). For example, for the graphemic lexicon:

```
ADAM'S          a^I  d^M a  ^M m  ^M;A s^F
```

Spontaneous speech contains hesitations and partial words. The former can take many forms such as 'er', 'um', 'hmmm' etc. For spoken language assessment the precise hesitation realisation is not important whereas detecting a hesitation has occurred is of interest, so for both types of lexicon all hesitations are mapped to the word %HESITATION%. To simplify the modelling of hesitations in the graphemic case, two further 'graphemes' are defined – G00, G01 – to act as two alternative pronunciations, capturing all hesitation variants (Gales et al 2015):

```
%HESITATION%    G00
%HESITATION%    G01
```

Partial words are modelled by pronouncing the part of the word that was spoken and marking that the word is not a complete word, for example:

Phonetic
```
DEG_%partial%    d eh g
DEG_%partial%    d ih g
DEG_%partial%    d iy g
```
Graphemic
```
DEG_%partial%    d  e  g
```

Crowd-source transcriptions

Unfortunately, the quantity of available data for non-native speech recognition, especially at the lower levels of proficiency, is very limited. For standard systems it is normally only necessary to collect data from fairly proficient users; learners (low-proficiency users) of a language are rarely the target users of speech-enabled systems (outside language learning and assessment applications). Thus, to get both acoustic and language model training data matched for the task of language learning and assessment standard, corpora cannot be used. The traditional approach for language assessment, however, where grades are generated manually, can be leveraged as a starting point. In particular, for some deployed systems such as the Business Language Testing Service (BULATS)[8] test, which is the task examined here, reasonable quantities of audio data and associated grades are available. The only element of data missing is the audio transcriptions required to train the acoustic and language models.

One of the challenges in obtaining these transcriptions is that human transcribers find it difficult to produce accurate transcriptions for low-proficiency speakers. In an initial study, a set of professional transcription services were asked to transcribe this data. The inter-annotator agreement was measured at approximately 25% word error rate (WER). This data was expensive to obtain, and took significant amounts of time. An alternative approach was adopted for these experiments: crowd-sourcing. Here multiple crowd-sourced (in this case Amazon Mechanical Turk) transcriptions were obtained and combined using an ASR system. This was cheaper, and significantly faster than using transcription services. An interesting question is how accurate the transcriptions are, and whether they are usable for ASR system building. To address this problem a small set of data, approximately 11 hours of audio, was carefully manually transcribed. This could then be treated as a *gold standard*.[9]

Table 1 shows the performance of the crowd-sourced transcriptions against the gold standard transcriptions. Note: due to the available data

8 Replaced with Linguaskill Business in 2019.
9 This ignores a range of issues associated with scoring speech recognition systems. For example, text normalisation does impact performance, e.g., O.K. and okay.

grades C1 and C2 are merged as C so that the number of candidates over the five grades is approximately equal.

Table 1 Crowd-sourced transcription % word error rates

Grade					
A1	**A2**	**B1**	**B2**	**C**	**Average**
28.2%	21.6%	15.4%	15.6%	9.2%	15.5%

Auto-marker

As shown in Figure 2, once the ASR system has been used to generate the word transcription, associated with the audio,[10] a set of features are extracted from the two sequences

$$\boldsymbol{z}^* = f(\boldsymbol{w}^*_{1:L}; \boldsymbol{x}^*_{1:T}) \tag{1.4}$$

The feature vector will be the same size for all responses, irrespective of the variable lengths of audio and text sequences observed. These features will be described in more detail in the next section. This set of features is then used by the grader module to evaluate the candidate's speech. Depending on the grader model selected, the set of features is used to estimate the proficiency of the candidate, which might be in the form of a discrete Common European Framework of Reference for Languages (CEFR, Council of Europe 2001) level or a continuous score corresponding to the CEFR proficiency scale (such as 0–6 for the Linguaskill Speaking test). Here, a Gaussian distribution over the proficiency of the candidates is obtained, from which we can extract information to provide feedback. The prediction of the score g extracted from the features $\boldsymbol{z}$ for a candidate can be expressed as

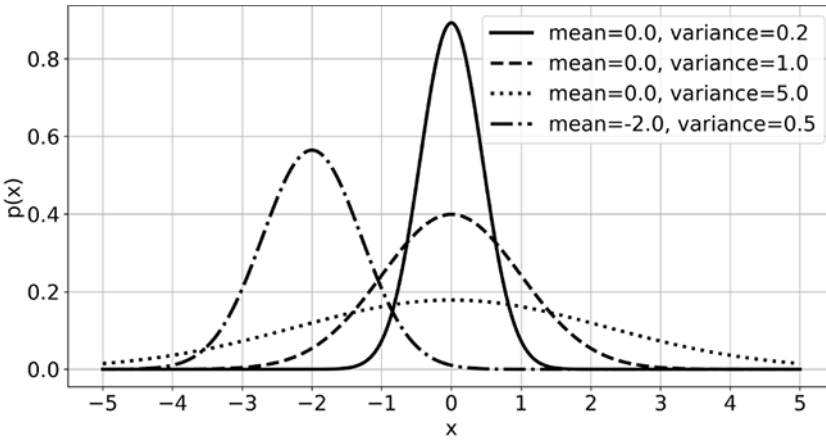
$$p(g^*|\boldsymbol{z}) = N(g^*; f_\mu(\boldsymbol{z}^*), f_\sigma(\boldsymbol{z}^*)) \tag{1.5}$$

For this chapter, a deep learning based approach is taken to predict the mean, $f_\mu(\boldsymbol{z})$, and variance, $f_\sigma(\boldsymbol{z})$, presented below.

The advantage of using a distributional representation is that in addition to the mean value representing the predicted score, the variance can, for example, be used to yield an estimate of how confident the auto-marker is in its prediction of a human score: a larger variance, σ^2, corresponding to a less precise fit to the mean as illustrated in Figure 7. This 'confidence' measure

10 Slight abuse of notation is used here as $\boldsymbol{x}$ is used to represent both the waveform and derived features. This is aimed to simplify the equations used.

Figure 7 Example of Gaussian (normal) distributions



allows the auto-marker to identify test entries that it cannot reliably score, such as a speaker with an unusual accent. This enables a hybrid or *human-in-the-loop* marking model such as the one used in the Linguaskill Speaking test (Xu et al 2020). Here, fast live automatic marking is applied, and then a number of confidence metrics produced by the auto-marker are used to determine if the test submission should be re-marked by human examiners. This ensures the quality of marking and provides human-marked data to further train the auto-marker. As the auto-marker improves with additional data, the number of human-examined tests will reduce.

A concern with any marking scheme is the scoring validity: 'the validity of the scores depends equally much on the rating criteria and the relationship between the criteria and the tasks' (Luoma 2004:171). The machine learning-based auto-marker learns from the labelled training data to predict similar scores to human examiners. Thus, the reliability of the auto-marker is directly related to the underlying manually judged assessments. Unlike human examiners, auto-markers do not get tired so have the potential to be more consistent than human examiners who might have different levels of experience and training, or simply be not at their best.

Auto-marker features

The auto-marker input features are selected to both optimise the performance of the auto-marker and 'to make sure that the features are reasonable representations of the construct of speech and that the scoring models are substantively meaningful' (Xi et al 2008:36). The features model a mixture of fluency, pronunciation, language resource, coherence/

discourse management and task achievement. The latter two, higher-order parts of the speaking construct are hard to model currently. Xi et al (2008) note that different parts of the speaking construct are highly correlated, so that not predicting these factors is not as detrimental to a model's agreement with human raters as it might be. For scoring validity, and to reduce the risk of gaming behaviours affecting the results, it is important that the features do not rely too heavily on a single aspect of proficiency (Loukina and Cahill 2016).

The auto-marker features can be split into three distinct groups depending on how they make use of the available information.

1. **Audio features**: derived directly from the speech signal without reference to what was said, such as the mean fundamental frequency and maximum energy. These *low-level signal processing* features (Higgins et al 2011) give a rough indication of the pitch and energy of the speaker but are not highly correlated with grade (Wang, Wong, Gales, Knill and Ragni 2018).
2. **Text features**: derived from the ASR hypothesis, these features attempt to give an indication of a learner's language resource. Vocabulary usage can be estimated using metrics such as the number of unique words. Use of grammar can be represented by features such as part-of-speech tag counts and the perplexity of the spoken test on grade-dependent language models (LMs) (Wang et al 2018). Higher-level learners will have richer vocabularies and be closer to higher-grade LMs than the text spoken by less proficient learners. These features can also be used in detecting an abnormal response (see the section 'Abnormal response detection').
3. **Audio and text features**: the majority of features are derived from a combination of text and audio. Features such as the mean phone duration, articulation rate and others related to stress and intonation contribute to pronunciation assessment; hesitation and rhythm-related features contribute to fluency assessment.

Deep learning-based auto-marker

As discussed in the ASR section, deep learning-based approaches have yielded state-of-the-art performance in a wide range of application areas. It is quite natural to apply these approaches to spoken language assessment. Deep learning is a highly flexible framework that involves multiple layers, hence the name deep, of highly connected layers of nodes. Each node performs a computation and has an *activation function* associated with it, which is normally non-linear. This yields a very powerful non-linear mapping process from the input to a set of targets. When applied to the grading this allows

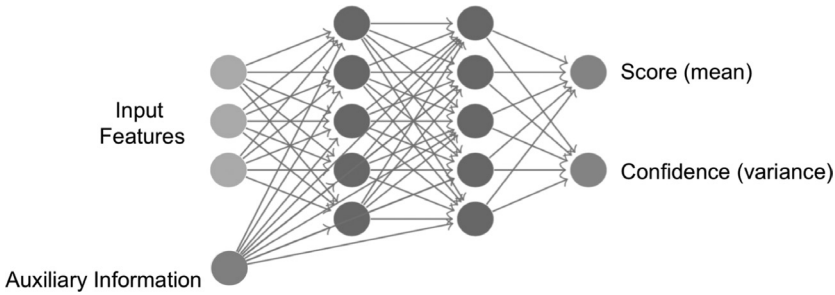
the output targets to be either CEFR grades (classification) or grade scores (regression).

Here, a regression approach is taken with the learner's score. A Deep Density Network (DDN) (Bishop 2006) is trained to predict a Gaussian distribution representation of the learner's score. DDNs are deep learning models consisting of a series of connected nodes with two outputs as shown in Figure 8. They are trained in a supervised learning fashion where the network is given many examples of training data in the form of pairs of input vectors and human-marker scores, $\{z_i, g_i\}$, and learns the mapping between the two. At test time the DDN infers the distribution for the input vector generated from the current learner's responses. The distribution mean, $f_\mu(z)$, is taken as the learner's score. The distribution variance, $f_\sigma(z)$, can be used to determine how uncertain the network is in its predicted score (Malinin, Ragni et al 2017). To further improve the reliability of the auto-marker score and the ability to predict when the system has made an incorrect prediction, an ensemble of DDNs can be trained (Wu, Knill, Gales and Malinin 2020). Multiple DDNs are trained with the same architecture but each with a different initialisation of the network parameters and the predictions from each combined, e.g., by averaging the scores. The variances across the ensemble also provide richer information about the uncertainty of the prediction. If computational cost is an issue, the ensemble models can be distilled into a single model which achieves close to the same performance.

Abnormal response detection

Learners may not provide expected responses to a question, for example, they might talk in their native language or might go off-topic by speaking a memorised general answer or speaking meaningless words. This might be a deliberate attempt to game the system or it could simply be that the learner does not understand the question or have the vocabulary to construct an appropriate response.

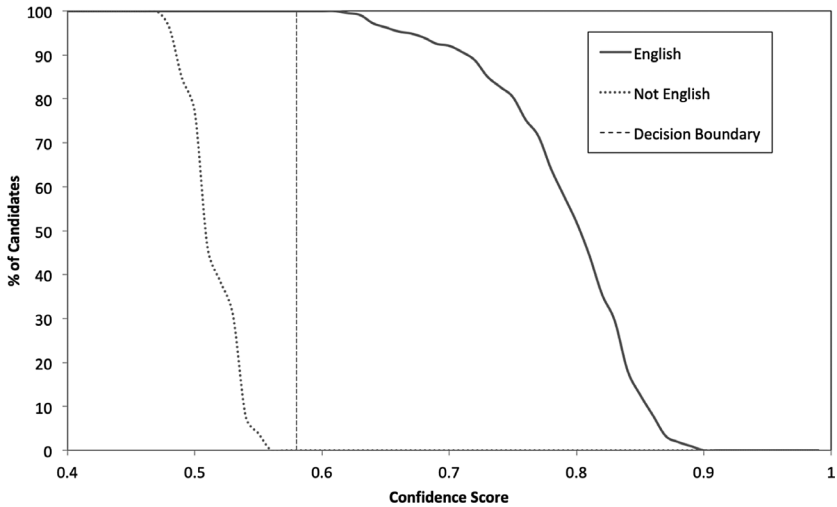
Figure 8 Deep Density Network (DDN)-based grader



One of the features of ASR systems is that they are only able to generate sequences of words that exist in the ASR vocabulary (the lexicon). Thus, even if a learner speaks in their first language, the ASR system will still generate word sequences in English.[11] Each word in the ASR hypothesis is accompanied by a confidence score which indicates how confident the ASR is that it has correctly hypothesised the word spoken. Mismatches between a predicted word and typical realisations of that word will yield lower confidence scores. Figure 9 shows the average confidence score from a range of learners either speaking their non-English L1, or English (see also Xu, Jones, Laxton and Galaczi 2021). It is clear from this plot that there is a clear decision boundary between the two sets of speakers, allowing learners not talking in English to be detected. In practice a learner might swap between English and their L1 depending, for example, on their vocabulary knowledge. Approaches for handling this form of *code-switching* is an active area of research.

A requirement for successfully completing a test is that the responses that a candidate gives are related to the question prompt. It is important to be able to detect *off-topic responses*. To give an example of this, consider the prompt word sequence $\boldsymbol{w}^{(p)}$:

Figure 9 Average speaker confidence scores for English and non-English (L1) speakers



11 Multilingual ASR foundation models can produce speech in other languages.

Prompt: Describe a company that you admire and why.

A valid response to this prompt, $\boldsymbol{w}^{(r)}$, might be[12]

Response: Cambridge University Press & Assessment is a wonderful company. It has a great working environment.

whereas an off-topic response might be

Response: Education is an important driver of social mobility.

From a machine learning perspective, the system needs to be able to generate the probability that for prompt $\boldsymbol{w}^{(p)}$ the response $\boldsymbol{w}^{(r)}$ is relevant, $P(\text{rel}|\boldsymbol{w}^{(p)}, \boldsymbol{w}^{(r)})$.

The issue of off-topic response detection has been studied in the context of essay (text) assessment (Landauer, Foltz and Laham 1998, Yannakoudakis 2013). More recently the area has been investigated for spoken responses using, for example, deep learning approaches to generate embeddings of the prompt and responses word sequences, and then using the relationship to determine relevance (Lee et al 2017, Malinin, Knill and Gales 2017, Malinin, Knill, Ragni, Wang and Gales 2017, Wang, Yoon, Evanini, Zechner and Qian 2019, Yoon et al 2017). Raina, Gales and Knill (2020) showed that these deep learning approaches are complementary and can be combined to boost detection of off-topic responses, particularly on responses to prompts that were not seen in training.

System performance

The auto-marking system described above has been assessed on data from the Cambridge Assessment English BULATS test that assesses English language skills for business, industry and commerce (Business Language Testing Service no date[13]). BULATS is a multi-level test with candidates from across the CEFR Levels A1–C2. The spoken part of the BULATS test consists of five parts:

- Section A: eight short prompt-response interview questions about the candidate and their background, work etc. (e.g., What is your job?)
- Section B: eight read-aloud sentences
- Section C: candidate talks for one minute on a business or work-related topic (e.g., the perfect office)

12 Cambridge Assessment funds the ALTA Institute.
13 Please note that since this study BULATS has been retired and replaced with Linguaskill Business.

- Section D: candidate talks for one minute about a graphic (e.g., pie chart or line graphs) with a business focus (e.g., sales figures)
- Section E: candidate answers five questions related to a business scenario (e.g., organising a conference); each answer is up to 20 seconds.

Data from BULATS tests was used to train and test the ASR and auto-marker. The data in its original form consists of speech waveforms, operational grades and metadata. Since the data comes from a deployed system, it includes a large variety of recording levels, background noises and other recording artefacts. The test set has 225 speakers from six L1s (Arabic, Dutch, French, Polish, Thai, Vietnamese), approximately equally distributed across L1s and the CEFR grades, with C1 and C2 grades merged into a single C grade due to lack of data. ASR word error rates are scored against manual transcriptions of the long free speaking sections, C–E, of the test set (Caines, Nicholls and Buttery 2017). Auto-marking is run over all five sections and a single score predicted for a candidate. It is assessed against grades produced by expert graders from Cambridge Assessment.

Speech recognition performance

A state-of-the-art deep learning-based non-native English ASR system is trained. The system is speaker-adaptive, i.e., it adapts the acoustic model to the speaker under test. It is a sequence teacher–student trained lattice-free maximum mutual information (LF-MMI) factorised time-delay neural network (TDNN-F) system (Povey et al 2016, 2018, Wang et al 2018, Wong and Gales 2016). 500 hours of data recorded under BULATS across a wide range of L1s and grades is used to train the acoustic models. The language models are an interpolation of models trained on transcriptions of the BULATS data with a general English language model. ASR lattices are generated with a trigram language model trained on 2.6m words. The lattices are rescored with a succeeding word RNNLM (su-RNNLM) (Chen et al 2017) trained on 25.6m words. The system is implemented using the Kaldi toolkit (Povey et al 2011) with filterbank features extracted using the HTK v3.5 toolkit (Young et al 2015) and su-RNNLM trained with the CUED RNNLM v1.1 toolkit (Chen, Liu, Gales and Woodland 2016). More details about the system can be found in Lu et al (2019) where it is named *ASR3*.

The overall WER is 19.5% (Table 2). This compares very favourably with inter-annotator word error rates of 25% seen on another subset of BULATS data transcribed using professional transcription services (Wang, Gales et al 2018). As the proficiency level decreases there is a corresponding rise in WER. Evaluation of A1 speech is particularly hard as transcribers often struggle to understand what is said at this level.

Table 2 Word error rates (% WER) for sections C–E

Grade					
A1	**A2**	**B1**	**B2**	**C**	**Average**
31.8%	25.4%	19.6%	18.0%	14.7%	19.5%

Auto-marker performance

The DDN grader was trained on examination data from approximately 1,000 candidates from the same L1s as the evaluation set.[14] Ten models were included in the DDN ensemble. Auto-marking features for both training and test were derived from ASR transcriptions, time-aligned at the word and grapheme level. The grader training set was held out from the ASR training data. Human examiners give each part of the BULATS test a holistic score on a scale from 0 to 6, marked at 0.5 increments to recognise higher-level speakers within a grade e.g., B1 has a score of 3 and B1-high 3.5. If the response contains no meaningful response or is off-topic, a score of 0 is assigned to the part, and the part is excluded from scoring if it cannot be marked due to, for example, audio quality issues. The overall score for a candidate is their average score across the five parts. For training, the target grades are taken from the original operational examiners in the field. The test data was re-marked by expert graders and these expert marks were used to evaluate the auto-marker predicted scores.

Pearson correlation coefficient (PCC) and mean square error (MSE) are used to measure intra/inter-rater reliability. As can be seen in Table 3, a strong correlation with expert grades is achieved. Furthermore, the predicted scores are relatively closely aligned to the expert graders with close to 94% of scores being within one grade point and 67% within half a point. These latter scores act as an approximation to the standard error of measurement (SEM) which can be applied to report reliability in terms of a confidence

Table 3 Auto-marking performance in terms of Pearson correlation coefficient (PCC) and mean square error (MSE), and the percentage of predicted scores within 0.5 and 1.0 points of the expert reference score on a BULATS test set

		Within	
PCC	**MSE**	**0.5**	**1.0**
0.884	0.320	67%	94%

14 There were no candidates in common between the ASR and grader test sets.

band around the examinee scores, although rarely reported for speaking assessment (Luoma 2004).

Conclusions

Automatic assessment of non-native spoken English is now a feasible proposition for open speaking tasks, across all levels of language proficiency. Machine learning-based approaches enable the auto-marking systems to learn from data to predict scores under the same construct as that used for human examiners. The use of deep learning and other advancements in ASR mean that the automatic transcription of the open speaking responses is at a level similar to that at which two human transcribers agree. Any ASR errors are mitigated in auto-marking by training on features derived from ASR transcriptions, leading to a good agreement between a DDN-based auto-marker and human examiners. This approach has been deployed in the Linguaskill Speaking test. Hybrid auto-marking (Xu et al 2020) is used to provide reliable assessments. When the auto-marker is confident in its scoring, the candidates receive the auto-marked score, otherwise their test submission is passed to a human examiner for marking. This has reduced the time and costs required for marking the Linguaskill speaking exams. A fully automated version is available for learners to access 24/7 in the Cambridge English Speak&Improve[15] research project with the University of Cambridge; a web-based system that helps people learning English to practise and measure their speaking ability. As more data is gathered and the technology developed, these systems will continue to improve in accuracy, consistency and reliability across L1s and proficiency levels. How to measure the performance of auto-marking systems in contrast to operational examiners is an ongoing area of research (Xu et al 2021).

One of the challenges when building assessment systems based on machine learning is the interpretability of the models. There is a range of ongoing research in this area that allows, for example, the importance of input features to be measured e.g.. their saliency (Simonyan, Vedaldi and Zisserman 2014), and analysis of the network layers e.g., through concept activation vectors (Kim et al 2018). These approaches however do not yield interpretations that are directly useful as a feedback mechanism to learners. Rather than doing more complex network interpretation and analysis it is possible to build systems that assess particular attributes (views) of a learner's English ability (Kyriakopoulos 2021). Here the holistic assessment score is generated from a series of scores for individual views such as intonation, pronunciation and content. Future developments in open speaking prompt-response tests will focus on aspects such as learning-orientated assessment,

15 speakandimprove.com

where feedback is provided to the learner about the areas they should focus on to improve, alongside their grade (Jones and Saville 2016). Multi-view assessment-based systems (Bannò, Balusu, Gales, Knill and Kyriakopoulos 2022, Kyriakopoulos 2021) will yield finer feedback detail. Research is still ongoing in areas such as handling children's speech, and measuring how well a learner has answered a question and carried out tasks such as summarisation and review. A lot of speaking exams are conversational in format, with a candidate talking to an interlocutor, and possibly with another candidate. Conversational (dialogic) assessment brings with it further challenges for automation (McKnight et al 2023). Apart from technical challenges in separating out individual speakers, the key challenge is how to measure communicative as well as linguistic speaking skills.

References

Bannò, S and Matassoni, M (2023) Proficiency assessment of L2 spoken English using wav2vec2.0, in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 1,088–1,095.

Bannò, S, Balusu, B, Gales, M, Knill, K and Kyriakopoulos, K (2022) View-Specific Assessment of L2 Spoken English, in *Proceedings of Interspeech 2022*, 4,471–4,475.

Bannò, S, Knill, K, Matassoni, M, Raina, V and Gales, M (2023) *Assessment of L2 Oral Proficiency Using Self-Supervised Speech Representation Learning*, 9th Workshop on Speech and Language Technology in Education (SLaTE), Dublin, August 2023.

Bernstein, J (1999) *PhonePass™ Testing: Structure and Construct*, Ordinate Corporation Technical Report, Menlo Park: Ordinate.

Bisani, M and Ney, H (2008) Joint-sequence models for grapheme-to-phoneme conversion, *Speech Communication* 50 (5), 434–451.

Bishop, C M (2006) *Pattern Recognition and Machine Learning*, Berlin: Springer Verlag.

BULATS. Business Language Testing Service (no date), available online: www.cambridgeenglish.org/exams-and-tests/bulats/

Caines, A, Nicholls, D and Buttery, P (2017) *Annotating errors and disfluencies in transcriptions of speech*, Technical Report UCAM-CL-TR-915, Cambridge: University of Cambridge, Computer Laboratory.

Chen, X, Liu, X, Gales, M and Woodland, P (2016) *CUED-RNNLM – an open-source toolkit for efficient training and evaluation of recurrent neural network language models*, available online: mi.eng.cam.ac.uk/projects/cued-rnnlm/papers/ICASSP16-Toolkit.pdf

Chen, X, Liu, X, Ragni, A, Wang, Y and Gales, M J F (2017) Future word contexts in neural network language models, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27 (9), 1,444–1,454.

Cheng, J, Chen, X and Metallinou, A (2015) Deep neural network acoustic models for spoken assessment applications, *Speech Communication* 73, 14–27.

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

Cucchiarini, C, Strik, H and Boves, L (1997) Automatic evaluation of Dutch pronunciation by using speech recognition technology, *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, 622–629.

Evanini, K and Wang, X (2013) Automated speech scoring for non-native middle school students with multiple task types, *INTERSPEECH 2013*, 2,435–2,439.

Franco, H, Abrash, V, Precoda, K, Bratt, H, Rao, R, Butzberger, J, Rossier, R and Cesari, F (2000) The SRI EduSpeak™ system: Recognition and pronunciation scoring for language learning, *Proceedings of InSTILL 2000*, 123–128.

Gales, M and Young, S J (2007) The application of hidden Markov models in speech recognition, *Foundations and Trends in Signal Processing* 1 (3), 195–304.

Gales, M J, Knill, K and Ragni, A (2015) Unicode-based graphemic systems for limited resource languages, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015*, 5,186–5,190.

Higgins, D, Xi, X, Zechner, K and Williamson, D (2011) A three-stage approach to the automated scoring of spontaneous spoken responses, *Computer Speech and Language* 25, 282– 306.

Hinton, G, Deng, L, Yu, D, Dahl, G E, Mohamed, A, Jaitly, N, Senior, A, Vanhoucke, V, Nguyen, P, Sainath, T N and Kingsbury, B (2012) Deep neural networks for acoustic modeling in speech recognition, *IEEE Signal Processing Magazine* 29 (6), 82–97.

Hu, W, Qian, Y, Soong, F and Wang, Y (2015) Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers, *Speech Communication* 67, 154–165.

Hu, Q, Richmond, K, Yamagishi, J and Latorre, J (2013) An experimental comparison of multiple vocoder types, in *Proceedings of 8th ISCA Workshop on Speech Synthesis (SSW 8)*, 135–140.

Jones, N and Saville, N (2016) *Learning Oriented Assessment: A Systemic Approach*, Studies in Language Testing Volume 45, Cambridge: UCLES/ Cambridge University Press.

Khabbazbashi, N, Xu, J and Galaczi, E D (2021) Opening the black box: Exploring automated speaking evaluation, in Lanteigne, B, Coombe, C and Brown, J D (Eds) *Challenges in Language Testing Around the World*, Springer, New York, 333–343.

Kim, B, Wattenberg, M, Gilmer, J, Cai, C J, Wexler, J, Viégas, F and Sayres, R (2018) Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV), in *Proceedings of International Conference on Machine Learning (ICML)*, available online: arxiv.org/pdf/1711.11279

Knill, K, Gales, M, Kyriakopoulos, K, Malinin, A, Ragni, A, Wang, Y and Caines, A (2018) Impact of ASR Performance on Free Speaking Language Assessment, in *Interspeech 2018*, 1,641–1,645.

Kyriakopoulos, K (2021) *Automatic Assessment of English as a Second Language*, PhD thesis, University of Cambridge.

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

Leacock, C and Chodorow, M (2003) C-rater: Scoring of short-answer questions, *Computers and the Humanities* 37 (4), 389–405.

Lee, C M, Yoon, S-Y, Wang, X, Mulholland, M, Choi, I and Evanini, K (2017) Off-topic spoken response detection using Siamese convolutional neural networks, in *Proceedings of Interspeech 2017*, 1,427–1,431.

Loukina, A and Cahill, A (2016) Automated scoring across different modalities, in *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, 130–135.

Loukina, A, Zechner, K, Chen, L and Heilman, M (2015) Feature selection for automated speech scoring, in *Proceedings of the Tenth Workshop on*

Innovative Use of NLP for Building Educational Applications (BEA), Denver: Association for Computational Linguistics, 12–19.

Lu, Y, Gales, M, Knill, K, Manakul, P, Wang, L and Wang, Y (2019) Impact of ASR performance on spoken grammatical error detection, in *Proceedings of Interspeech 2019*, 1,876–1,880.

Luoma, S (2004) *Assessing Speaking*, Cambridge: Cambridge University Press.

Ma, R, Qian, M, Gales, M and Knill, K (2023) *Adapting an ASR Foundation Model for Spoken Language Assessment*, 9th Workshop on Speech and Language Technology in Education (SLaTE), Dublin, August 2023.

Malinin, A, Knill, K and Gales, M (2017) Hierarchical attention based model for off-topic spontaneous spoken response detection, in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) 2017*, 397–403.

Malinin, A, Knill, K, Ragni, A, Wang, Y and Gales, M (2017) An attention based model for off-topic spontaneous spoken response detection: An initial study, in *Proceedings of ISCA Workshop on Speech and Language Technology for Education (SLaTE)*, 144–149.

Malinin, A, Ragni, A, Knill, K and Gales, M (2017) Incorporating Uncertainty into Deep Learning for Spoken Language Assessment, in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver: Association for Computational Linguistics, 45–50.

McKnight, S W, Civelekoglu, A, Gales, M, Bannò, S, Liusie, A and Knill, K M (2023) *Automatic Assessment of Conversational Speaking Tests*, 9th Workshop on Speech and Language Technology in Education (SLaTE), Dublin, August 2023.

Metallinou, A and Cheng, J (2014) Using deep neural networks to improve proficiency assessment for children English language learners, in *Interspeech 2014*, 468–1472.

Mikolov, T, Karafiát, M, Burget, L, Černocký, J and Khudanpur, S (2010) Recurrent neural network based language model, in *Proceedings of Interspeech 2010*, 1,045–1,048.

Nicholls, D, Knill, K M, Gales, M J F, Ragni, A and Ricketts, P (2023) Speak & Improve: L2 English Speaking Practice Tool, in *Proceedings of Interspeech 2023*, 3,669–3,670.

Novak, J, Minematsu, N and Hirose, K (2015) Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework, *Natural Language Engineering* 22 (6), 907–938.

Povey, D, Ghoshal, A, Boulianne, G, Burget, L, Glembek, O, Goel, N, Hannemann, M, Motlíček, P, Qian, Y, Schwarz, P, Silovský, J, Stemmer, G and Veselý, K (2011) The Kaldi speech recognition toolkit, in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) 2011*, available online: www.danielpovey.com/files/2011_asru_kaldi.pdf

Povey, D, Cheng, G, Wang, Y, Li, K, Xu, H, Yarmohamadi, M and Khudanpur, S (2018) Semi-orthogonal low-rank matrix factorization for deep neural networks, in *Proceedings of Interspeech 2018*, 3,743–3,747.

Povey, D, Peddiniti, V, Galvez, D, Ghahrmani, P, Manohar, V, Na, X, Wang, Y and Khudanpur, S (2016) Purely sequence-trained neural networks for ASR based on lattice-free MMI, in *Proceedings of Interspeech 2016*, 2,751–2,755.

Qian, Y, Lange, P, Evanini, K, Pugh, R, Ubale, R, Mulholland, M and Wang, X (2019) Neural approaches to automated speech scoring of monologue

and dialogue responses, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2019*, 8,112–8,116.

Raina, V, Gales, M and Knill, K (2020) Complementary systems for off-topic spoken response detection, in *Proceedings of Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, 41– 51.

Richmond, K, Clark, R A J and Fitt, S (2009) Robust LTS rules with the Combilex speech technology lexicon, in *Proceedings of INTERSPEECH 2009*, 1,295–1,298.

Simonyan, K, Vedaldi, A and Zisserman, A (2014) Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, in *Proceedings of the Workshop at International Conference on Learning Representations (ICLR)*, available online: arxiv.org/pdf/1312.6034

The CMU Pronouncing Dictionary (no date) available online: www.speech.cs.cmu.edu/cgi-bin/cmudict

Tao, J, Ghaffarzadegan, S, Chen, L and Zechner, K (2015) Exploring deep learning architectures for automatically grading non-native spontaneous speech, in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6,140–6,144.

van Dalen, R, Knill, K and Gales, M (2015) Automatically grading learners' English using a Gaussian Process, in *Proceedings of ISCA Workshop on Speech and Language Technology for Education (SLaTE)*, available online: www.vandalen.uk/pdf/van_dalen-2015-grading.pdf

Wang, X, Yoon, S-Y, Evanini, K, Zechner, K and Qian, Y (2019), Automatic Detection of Off-Topic Spoken Responses Using Very Deep Convolutional Neural Networks, in *Proceedings of Interspeech 2019*, 4,200–4,204.

Wang, Y, Wong, J H M, Gales, M, Knill, K and Ragni, A (2018) Sequence teacher-student training of acoustic models for automatic free speaking language assessment, in *Proceedings of IEEE Spoken Language Technology Workshop (SLT) 2018*, 994–1,000.

Wang, Y, Gales, M J F, Knill, K, Kyriakopoulos, K, Malinin, A, van Dalen, R C and Rashid, M (2018) Towards automatic assessment of spontaneous spoken English, *Speech Communication* 104, 47–56.

Witt, S M (1999) *Use of speech recognition in computer-assisted language learning*, PhD thesis, University of Cambridge.

Wong, J H M and Gales, M (2016) Sequence student-teacher training of deep neural networks, in *Proceedings of Interspeech 2016*, 2,761–2,765.

Wu, X, Knill, K, Gales, M and Malinin, A (2020) Ensemble Approaches for Uncertainty in Spoken Language Assessment, in *Proceedings of Interspeech 2020*, 3,860–3,864.

Xi, X, Higgins, D, Zechner, K and Williamson, D M (2008) *Automated Scoring of Spontaneous Speech Using SpeechRater v1.0*, Technical Report ETS RR-08-62, Princeton: Educational Testing Service.

Xiong, W, Droppo, J, Huang, X, Seide, F, Seltzer, M L, Stolcke, A, Yu, D and Zweig, G (2017) Toward human parity in conversational speech recognition, *IEEE/ACM Transactions on Audio, Speech and Language Processing* 25 (12), 2,410–2,423.

Xu, J, Jones, E, Laxton, V and Galaczi, E (2021) Assessing L2 English speaking using automated scoring technology: examining automarker reliability, *Assessment in Education: Principles, Policy & Practice* 28 (4), 411–436.

Xu, J, Brenchley, M, Jones, E, Pinnington, A, Benjamin, T, Knill, K, Seal-Coon, G, Robinson, M and Geranpayeh, A (2020) *Linguaskill – building a*

validity argument for the Speaking test, Cambridge: Cambridge Assessment English.

Yannakoudakis, H (2013) *Automated assessment of English-learner writing*, Technical Report UCAM-CL-TR-842, Cambridge: University of Cambridge, Computer Laboratory.

Yoon, S-Y, Lee, C M, Choi, I, Wang, X, Mulholland, M and Evanini, K (2017) Off-topic spoken response detection with word embeddings, in *Proceedings of Interspeech 2017*, 2,754–2,758.

Young, S, Evermann, G, Gales, M, Hain, T, Kershaw, D, Liu, X, Moore, G, Odell, J, Ollason, D, Povey, D, Ragni, A, Valtchev, V, Woodland, P C and Zhang, C (2015) *The HTK book (for HTK version 3.5)*, Cambridge: University of Cambridge.

Yu, D and Li, J (2017) Recent progresses in deep learning based acoustic models, *IEEE/CAA Journal of Automatica Sinica* 4, available online: arxiv.org/pdf/1804.09298

Zechner, K, Higgins, D, Xi, X and Williamson, D M (2009) Automatic scoring of non-native spontaneous speech in tests of spoken English, *Speech Communication* 51 (10), 883–895.

14 Using technology and statistics to detect cheating in objectively marked tests

Edmund Jones
Cambridge University Press & Assessment, UK

Abstract

Cheating is a threat to test validity and happens by a wide variety of means. This chapter is about statistical and machine learning methods to detect cheating after the fact, in multiple-choice tests and other objectively marked tests, based on the test-takers' responses and possibly other variables such as response times for computer-based tests. Modern statistical methods for detection of cheating generally use item response theory, which is a set of psychometric models for whether test-takers give the correct responses to items. These methods are well established and understood. In recent years, research has begun on the use of machine learning to detect cheating. Some machine learning methods are interpretable and others are more like a black box. The chapter provides an introduction to these topics and a discussion of possible future directions of research.

Introduction

Cheating on tests is a perennial and important problem. It can be defined as misconduct by test-takers, administrators, teachers, or other persons, that is intended to result in test-takers getting higher scores than they properly should (Cizek 1999). In the UK the term 'malpractice' is often used with this meaning. Cheating has been widely discussed in the field of general educational assessment, but barely at all in the language testing literature. One article about the fairness of language tests made no mention of the issue (Xi 2010).

In all types of assessment, technology has changed the way test-takers cheat and the way testing organisations prevent and detect cheating in recent years. This chapter is primarily about one type of technology: the statistical and computational methods used by testing organisations to detect cheating

after the event, for multiple-choice tests as used for listening and reading. Almost everything in the chapter also applies to other objectively marked items such as matching tasks and single-word cloze items. There are several reasons for focusing on these types of tests. Firstly, cheating on them is easier and probably more prevalent than cheating on other types of tests, since the test-taker only needs to fill in a single bubble or write a single word. Secondly, there has been much research on cheating on these tests because of their prevalence in the US in the post-war years (Cizek 1999:37). Thirdly, cheating on these tests is more susceptible to detection by computational methods.

The following section is a discussion of cheating on tests in general, including examples of major incidents and a summary of how testing organisations deter and prevent cheating before it happens. The next two sections focus on cheating on language tests in particular and how it is related to the different types of test validity, and statistical methods that are used to detect cheating. The chapter then covers the machine learning methods that researchers have recently started to investigate, and cluster analysis, a particular class of machine learning methods. The final two sections provide possible avenues for future research and a conclusion.

How cheating is committed and prevented

To understand the general phenomenon of cheating, it is useful to look at several stories that have reached the news.

In the Atlanta Public Schools (APS) scandal, around 2008–11, there was widespread misconduct on a multiple-choice examination given annually at all public schools in the US state of Georgia (Kingston 2013). Staff erased students' incorrect answers after the tests and replaced them with correct answers, in response to unrealistic targets set by the central administration and 'unreasonable pressure' to achieve them (Bowers, Wilson and Hyde 2011:350). Investigators appointed by the state governor identified 178 educators who had been involved in cheating (Bowers et al 2011). The main physical evidence in the APS scandal was the answer sheets with incorrect answers erased and correct answers pencilled in. Scanners were able to detect the faint grey marks that remain after pencilled answers are erased.

In 2016 Reuters reported on widespread cheating on the SAT test, used for admissions to US colleges (Dudley, Stecklow, Harney and Liu 2016). Questions used in tests in the US were routinely reused in later tests that took place overseas; one student reportedly saw the exact same test in March 2015 in the US and October 2015 in Europe. Test preparation schools in East Asia were thus able to gather material from the tests and provide it to their customers as a study aid. This practice was admitted to by the manager of

one school in Shanghai, China that had 8,000 students each year. Reuters found eight instances in 2013–16 where material from a test was available online before the test was administered. Estimates of the total numbers of students who might have benefitted from this kind of cheating were not reported, but the SAT was taken by 64,000 students in East Asia in the 2013–14 academic year.

In 2019 there was a major scandal related to cheating on college admissions in the US. Test centre administrators were bribed to allow a different person to take the test in place of the actual student, or to correct the students' answers after the test. Several dozen people were charged with federal crimes (BBC 2019, United States Department of Justice 2019a, 2019b).

In the UK, there was a cheating scandal related to TOEIC (the Test of English for International Communication). In February 2014, the TV programme *Panorama* exposed organised cheating on this test, which was one of the officially approved English language tests used by students and other immigrants to fulfil visa requirements (BBC 2014). In one scene filmed by a hidden camera, test-takers for a computer-based speaking test entered the room but were then told to step aside, and the tests were taken by different people in place of them (this type of cheating is often difficult or impossible to detect by statistical methods like the ones discussed in this chapter). In another, the invigilator (proctor) read out all the answers to a multiple-choice test. The cheating was organised and test-takers paid money to a company in London for the service.

The Home Office subsequently suspended test centres' licences and requested Educational Testing Service (ETS) to stop running TOEIC tests for UK immigration. ETS reviewed tens of thousands of speaking tests, using voice recognition software and human listeners, and judged that there had been cheating in 58% of cases (National Audit Office 2019). For two years, the Home Office revoked visas in all such cases. The story then took a different turn. Many test-takers claimed they had been wrongly accused, and appealed to the courts, and the Home Office was accused of wrongly revoking visas based on weak evidence (Bindmans 2016, Gentleman 2019, Main 2016).

In general, cheating on tests happens all around the world and by many means. Test-takers peek at the answers of the person next to them, smuggle illicit notes or gadgets into the room, smuggle notes or test-papers out of the room, write information on their arms or clothes, or use innocuous-looking signals to pass information (Cizek 1999, Geranpayeh 2013). Cheating behaviours have changed over time as technology evolves. Mobile phones have been used since the early days of their mass popularity (Maslen 1996). Tiny radio devices and cameras have been available for decades (Cizek 1999:55–58), and recently it has become much easier for cheaters to steal and share test content by using tiny cameras that are permanently connected to the internet (Maynes 2018). Foster (2013) described two gadgets designed

for stealing test items: ButtonCam, a camera disguised as a shirt button, and DocuPen, which looked like a normal pen but was able to record while being swiped across the page or screen.

Naturally, testing organisations use a range of measures to deter and prevent cheating, starting with proper test administration, trained invigilators, and checks on test-takers' identities (Wollack and Fremer 2013b). These measures have evolved in recent years in parallel with the changes in ways of cheating. Wollack (2018) mentioned biometric measures to verify test-takers' identities, high-definition video cameras that make it possible to observe test-takers from multiple angles, and software that prevents people taking computer-based tests from using other software such as web browsers. At least one organisation runs computer-based language tests in which the camera on the computer is used to routinely photograph the test-taker during the test (iTEP 2018). Foster, Mattoon and Shearer (2009) described a study using a high-security system where test-takers were required to purchase a specific webcam and set it up so their face, upper body, and keyboard could be viewed by remote invigilators.

When computer-based adaptive tests were introduced, it was believed that they would make cheating much more difficult, partly because test-takers do not usually see the same questions as the people next to them (Cohen and Wollack 2006, Wollack and Fremer 2013b). But such tests involve items appearing at different times in different test sessions, and this entails other risks, such as a test-taker using a concealed video camera to record all the items that appear on the screen and then posting them on the web. A group of test-takers might even be able to record most of the items in a test bank, if they arranged in advance to give responses that would lead to a wide range of items being shown.

The issue of cheating has been taken more and more seriously by major educational associations in recent years. An example of this can be seen in the two most recent editions of the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME] and Joint Committee on Standards for Educational and Psychological Testing 1999, 2014). The 1999 edition contained several standards that referred to cheating, whereas the 2014 edition included a new and more explicit standard for testing organisations that stated, 'reasonable efforts should be made to ensure the integrity of test scores by eliminating opportunities … to obtain scores by fraudulent or deceptive means' (2014:116). The NCME has published a policy paper on testing and data integrity (Bishop et al 2012) and a position statement on test security (NCME 2019). The Conference on Test Security has been running annually since 2012 and the Association of Test Publishers (ATP) held its first Global Security Summit in 2020. ATP also has a standing committee on test security. In the UK, a commission was launched

in 2018 to investigate cheating in the national examinations for school and vocational students and protect the integrity of the system against technological threats (Hazell 2019, Joint Council for Qualifications 2018).

Cheating on language tests

Cheating on language tests happens in broadly the same ways as cheating on other types of tests, and it is combated in broadly the same kinds of ways. Responses to multiple-choice tests can be analysed in the same way whatever the type of test. However, when speaking or writing tests are considered, there are some issues that are particular to language testing. Speaking tests are presumably difficult to cheat on, because the examiner is sitting in front of the test-taker and the test-taker has to physically pronounce words and phrases. But a test-taker might be able to cheat by acquiring the questions or prompts in advance and memorising suitable answers, or employing someone to impersonate them. Writing tests, especially higher-level tests that require the test-taker to write long passages, are probably harder to cheat on than multiple-choice tests. But test-takers' capacity for memorisation should not be underestimated (Watkins and Biggs 1996).

This volume is about validity in language tests. Cheating is obviously a threat to the validity of test scores, since if people cheat successfully, their scores will not accurately reflect their levels of ability. In a norm-referenced test, the scores will be inaccurate even for the test-takers who did not cheat. In a criterion-referenced test, the scores should be accurate for the honest test-takers, but their interpretation might not be, since cheating by others will change the rank-ordering of test-takers; and if cheaters' responses had been used in item calibrations, then all scores would be inaccurate.

In terms of Weir's socio-cognitive framework for the validity of language tests (Weir 2005, Weir and Shaw 2005), cheating is objectionable in terms of several types of test validity. Firstly, it diminishes the cognitive validity of a test, since the cognitive processes involved in passing illicit messages or bribing an invigilator are completely different from the cognitive processes that the test is designed to measure. Secondly, cheating is related to consequential validity, which is the effects that test scores have on test-takers and other parties. If a person cheats on a language test in order to work as a doctor or a pilot, the consequences for their patients or passengers could be severe. Consequential validity also includes the fairness of a test, which is obviously weakened by cheating. Thirdly, cheating affects context validity, which includes the administrative conditions of the test. Some test security measures diminish the authenticity of the situation – when writing a piece of English in normal daily life, you do not have to surrender your mobile phone and sit in silence 1.25 metres away from the next person – but these are clearly necessary to prevent cheating.

Cheating is, of course, a misdeed that needs to be combated in the interests of justice and fairness. Nevertheless, in some places it is viewed as culturally acceptable, and in some educational systems it is rampant (Noah and Eckstein 2001:36, 40). One advantage of regarding cheating as an issue of validity is that this circumvents any debates about cultural differences or moral relativism, to some extent (Albanese and Wollack 2013, Cizek and Wollack 2017a, Kim, Woo and Dickison 2017).

Statistical methods for detecting cheating

This section is about statistical methods that are used to detect anomalous patterns in test-takers' responses to multiple-choice tests. If a test-taker is flagged by one of these methods as a possible cheater, further evidence can be sought and a judgement can be made about whether to cancel the score and whether any other measures should be taken. These methods can detect anomalies caused by a wide variety of cheating behaviours, including a test-taker peeking at the answer sheet of the person next to them, or acquiring the answers in advance from a website, or centre-scale cheating as when the TOEIC invigilator read out the answers to the whole room. It is worth mentioning that anomalous patterns can arise from causes other than cheating, so these statistical methods are often described by saying that they flag responses or test-takers as possibly aberrant, rather than as cases of possible cheating; and statistical methods, by their nature, do not provide absolute proof that cheating or anything else abnormal happened.

The use of statistical methods to detect cheating dates back to Bird (1927, 1929), around the time when the basic ideas of statistical science were being developed. Several publications appeared in the subsequent decades, for example Saupe (1960) and Angoff (1974). The field grew rapidly from the late 1990s and there are now dozens of methods available, of which some are very well established (Cizek and Wollack (Eds) 2017b, Wollack and Schoenig 2018). The increase in the use of statistical methods was of course largely due to the great increases in the prevalence and power of computers.

One important class of statistical methods is person-fit indices, which analyse an individual test-taker's responses, for example to see whether they got a certain set of difficult items right despite having low ability according to other items (Karabatsos 2003, Zopluoglu 2017). Modern person-fit indices are mostly derived from statistical models such as item response theory (IRT), a class of models for test responses that include parameters for test-takers' abilities and item difficulties. These indices can help in the detection of a variety of cheating behaviours, such as advance knowledge of items and tampering with the response sheet after the test, but they do not perform very well for detecting copying, for example where the test-taker leaned over and peeked at their neighbour's answers (Zopluoglu 2017).

Another large class of indices is based on the responses from a pair of test-takers. Directional copying indices are designed to detect whether a specific one of the pair copied from the other (or the pair cheated in some other copying-like way), whereas symmetrical similarity indices treat the two test-takers in the same way and are designed to detect broader types of collusion such as the test-takers both receiving illicit assistance from the same person. Some of these pairwise indices are based on identical wrong answers – the idea being that if two test-takers give the same right answer then that is probably innocent, but if they give the same wrong answer then this is more suspicious. Others are based on all answers that are the same, whether they are right or wrong – the idea being that both of these might be due to cheating. Copying and similarity indices perform better than person-fit indices for the detection of copying (Zopluoglu 2017), and among the best-performing indices are omega (Wollack 1997) and M4 (Maynes 2014, 2017).

With modern computers, person-fit indices for all test-takers can be calculated very quickly. Calculating a similarity index for screening all test-takers is more time-consuming because the number of pairs is much larger – if there are 50,000 test-takers there are almost 1.25 billion pairs. But number-crunching tasks like this can be performed quickly if large numbers of computer processors are used in parallel. All these indices can also be used in a confirmatory way, for example if an invigilator reports suspicious behaviour by a group. The number of test-takers to be investigated will usually then be much smaller and it will be easy to calculate similarity indices for all pairs.

Another class of methods uses the response times in computer-based tests. For example, if a test-taker answers most items in an average way but answers a certain set of difficult items very quickly and correctly, then it might be suspected that they are cheating (Qian, Staniewska, Reckase and Woo 2016). Van der Linden (2007) presented a framework for modelling responses and response times that enables realistic modelling of test-takers' speed and accuracy. Response times can be useful because they are continuous variables (unlike item responses), and cheaters are unlikely to be able to avoid detection by deliberately modifying their response times. However, one recent study achieved only moderate success when using two response time methods on a real dataset (Boughton, Smith and Ren 2017).

In the set of methods known as score differencing, a test-taker's performance on one test, or part of a test, is compared to their performance on a different test, or part of a test (Bishop and Egan 2017). For example, it might be suspicious if a person took the same test twice within a short space of time and their performance improved greatly (though the way they spent the intervening time needs to be taken into account), or if their performance differed greatly between separate parts of the test that were supposed to have similar difficulty. For example, consider the Common European Framework

of Reference for Languages (CEFR, Council of Europe 2001), in which there are six levels ranging from A1 ('beginner') to C2 ('proficient'). If a test-taker achieved C2 in writing but only B1 in reading then their writing score would be regarded as fishy. Imbalances like this could be investigated using the combinations of scores achieved by other test-takers. The higher the correlation between the score of interest and the other score, the more useful the information is.

Score differencing methods can also be used to compare a test-taker's scores on items that are believed to have been leaked on the internet, or compromised in some other way, with items that are believed to be secure. Research on score differencing first appeared in the early 2000s and includes simple and sophisticated models. For school- or district-level anomalies, simple methods might be preferred because, all other things being equal, they are easier to explain to the media and the public.

Most methods discussed in this section use statistical hypothesis tests. In many fields of social and natural science, hypothesis tests are usually performed so that there is a 5% chance of a false positive – for example, there is only a 5% chance that a new psychological test will be wrongly judged to be more accurate than the current one. However, in cheating detection, testing organisations usually take the position that false positives are extremely undesirable, because accusing someone of cheating on a test is a serious matter and it is important to avoid false accusations. Therefore the threshold is usually set much lower (Skorupski and Wainer 2017, Zopluoglu 2017). The exact value has to be decided by each testing organisation (Maynes 2017).

If a hypothesis test is used many times to screen a large number of test-takers, then it is essential to deal with the issue of multiple testing. Suppose a hypothesis test on a person-fit index is carried out on all test-takers with the chance of a false positive set to 0.1%. If the issue of multiple testing is ignored, then it can be expected that 0.1% of test-takers will be spuriously flagged as possible cheaters. The most common way to deal with multiple testing is the Bonferroni correction (Dunn 1961, Wasserman 2004:166), but other approaches are available (Benjamini and Hochberg 1995, Benjamini and Yekutieli 2001).

When a similarity or copying index is used on all pairs of test-takers, the situation is more complicated, mainly because each test-taker is now involved in a large number of hypothesis tests which are not statistically independent. If the index is only meant to detect copying from nearby test-takers, then only pairs who sat close to each other need to be analysed, and this must be taken into account when adjusting for the multiple testing. For further information on multiple testing in relation to cheating detection, see Wollack, Cohen and Serlin (2001) and Maynes (2017).

Machine learning

Machine learning (ML) has started to be used for the detection of cheating on tests in recent years. ML is similar to the modern field of statistics but evolved from computer science rather than mathematics. ML methods for cheating detection are still unproven – at the time of writing they have only been evaluated in one study each – and are not as well understood or widely used as the statistical methods in the previous section, so it is not yet possible to make strong statements about how well they perform.

Many ML methods are supervised – they focus on prediction, which means estimating the value of one variable given the values of other variables. For example, the responses to a test and the response times could be used to predict whether a test-taker cheated. Supervised methods can be evaluated by using them on new data in which all the variables are known. Some ML methods are unsupervised, which means the goal is to infer properties of the distribution of the variables (or some similar task). Cluster analysis, discussed in detail in the next section, is unsupervised. When used on their own, unsupervised methods are difficult to evaluate since there is no obvious criterion for whether they are performing well (Färber et al 2010, Hastie, Tibshirani and Friedman 2009:487).

Thomas (2016) investigated the use of support vector machines (SVMs) to detect whether items had been compromised – that is, seen by test-takers before the test. The input variables she used included parameter estimates from IRT models, the Rasch infit and outfit statistics (which measure the differences between a test-taker's expected and observed responses), average response times, and other variables. Thomas used the method on data from a healthcare certification test for which the test publisher had discovered that certain items might have been compromised.

Man, Harring and Sinharay (2019) compared ML methods with traditional statistical methods for the detection of cheating. They used three unsupervised methods and three supervised methods. For the unsupervised methods they told the algorithm to classify the test-takers into precisely two groups, and after running the algorithm they decided which group was the cheaters by looking at variables that differed between the two groups. This was not a fully specified or automated procedure to flag test-takers for possible cheating.

Kim et al (2017) used two datasets in which certain test-takers were suspected or known to have cheated (Cizek and Wollack 2017a), and set out to find characteristics that tended to be shared by those test-takers. This was different from actually trying to detect cheating and seems to have been more like an early-stage exploratory method to provide clues for further investigation. They used an unsupervised ML method called market basket

analysis, which is able to deal with very large datasets and large numbers of variables (Hastie et al 2009:488–489).

Cluster analysis

This section is a more detailed discussion of one class of computational methods, known as cluster analysis. Cluster analysis falls squarely within both ML and statistics, and it has recently been used in a handful of publications about cheating detection. Like the ML methods in the previous section, cluster analysis is new in the field of cheating detection and has not yet been widely studied or become well established.

Clustering simply means putting objects into groups. Applications include a company grouping its customers according to demographic factors and the products they buy, a scientist grouping tumours by their genetic sequences, or a botanist grouping flowers according to their dimensions or shapes.

Cluster analysis for detecting cheating

When cluster analysis is used to investigate cheating on tests, the most notable type of cluster is likely to result from the existence of a leaked answer key (or some other kind of 'latent source'). The test-takers will have copied their responses from this key and so their responses will be more similar than if there had not been any cheating. But the test-takers will not all give exactly the same responses. They might make mistakes when copying, or choose different answers because they think they know better than the source. They might even choose different answers in an attempt to avoid detection. The leaked answer key might not include all the items on the test.

This kind of cheating can be hard to detect because the existence and content of the leaked answer key are usually not known. If they are known, there is no need for cluster analysis and test-takers can simply be flagged if their responses are very similar to the leaked answer key (Scott, Cooper and Maynes 2015).

Cluster analysis can be seen as an evolution of the statistical methods described above. Many of those methods come from a pre-digital era when a testing organisation could safely assume that if two or more test-takers cheated together, they must have been sitting close to each other during the test, or at least been in the same examination centre. Screening or investigations only needed to consider a single centre at a time. These days such assumptions can no longer be made, and there might be a group of test-takers who have colluded despite being scattered across different centres. For example, the answer key might have been leaked on the internet. The idea of using cluster analysis is to identify groups like these.

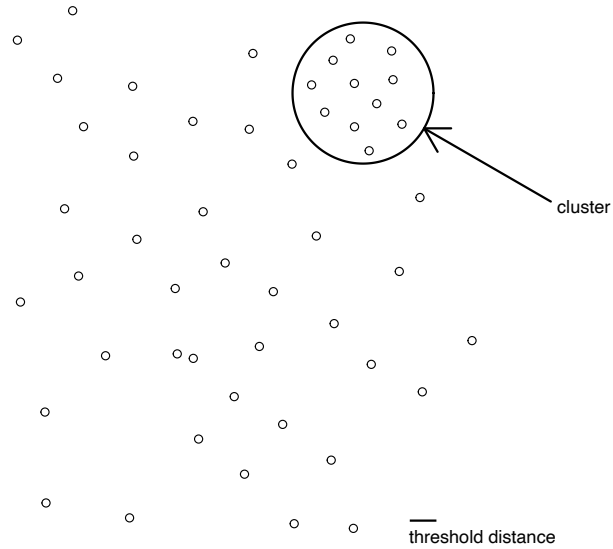
This is illustrated in an intuitive way by Figure 1. Suppose that each circle represents a test-taker, and the distance between two circles represents some measure of the 'distance' between the two test-takers' sets of responses (other variables can also be used). A traditional statistical method using the threshold shown at the bottom right would flag a pair as possible cheaters if they were closer to each other than this threshold. In this case only the pair slightly below and left of the middle would be identified. The group of test-takers in the top-right of the figure look suspiciously close together, but none of their pairwise distances are below the threshold, so the traditional method will not identify them. Cluster analysis ought to be able to automatically find this group of suspicious test-takers.

If the clustering method uses a threshold, this needs to take multiple testing into account – false positives are even worse if a whole cluster is wrongly accused – and this can be done in several ways depending on the clustering method (Belov and Wollack 2018, Wollack and Maynes 2017).

There is a wide variety of methods to perform the clustering (Hennig and Meila 2016). One important question is what measure of distance to use. The distance between two test-takers has to be a measure of the similarity of their two sets of variables (higher distance means lower similarity). If the variables are the responses, the distance could be a similarity index such as omega (Wollack 1997) or M4 (Maynes 2014, 2017). Other possibilities are the Hamming distance, which is essentially the number of items for which the test-takers' responses differ, the Euclidean distance, and the cosine similarity (Contreras and Murtagh 2016).

For multiple-choice items, it is necessary for the responses to be converted to a numerical form, so each possible response to each item is stored as a binary variable. For example, if an item has possible responses A, B, C, and D then this requires four binary variables, and if a test-taker chooses C then these take the values 0, 0, 1, and 0. For binary data there are at least 39 possible distance metrics (Shi 1993).

Figure 1 A set of test-takers including a possible cluster of cheaters in the top right. A traditional pairwise method using the threshold shown would not identify this cluster; cluster analysis might work better.



Research using cluster analysis

The first publication on the use of cluster analysis to detect cheating was Wollack and Maynes (2017). The clustering method they used is known as agglomerative hierarchical clustering with single linkage, and the distance metric was the M4 index (Maynes 2014). Wollack and Maynes used a fixed threshold, calculated using a multiple-testing correction at the test-taker level so that the probability of wrongly flagging a test-taker was 5%. All of this means that a pair of test-takers were placed in the same cluster if its M4 value exceeded 1.301, or if there was a chain of test-takers between them with all the M4 values over 1.301. This method is straightforward and objective, though cluster analysis is not usually performed in a pre-determined way like this.

Secondly, Maynes and Thomas (2017) described a two-stage method: first use the clustering method of Wollack and Maynes (2017), then take the largest cluster and estimate a latent source (leaked answer key) for that cluster. The main purpose of this research was to compare different methods for the second stage.

Thirdly, as mentioned in the section 'Machine learning', Man et al (2019) attempted to identify cheaters by using k-means clustering (Hartigan and Wong 1979, Mirkin 2016), among other methods. In k-means, objects are

grouped into a pre-specified number of clusters, k. Each cluster has a centre and the algorithm minimises the sum of the squared distances from each point to the centre of its cluster (Kriegel, Schubert and Zimek 2017). Advantages of k-means are that it is fast and many software packages can perform it; disadvantages are that k has to be chosen in advance and the algorithm only works with Euclidean distance. Man et al used $k = 2$, which removes the difficulty of choosing k but gives rise to two other possible problems. Firstly, k-means tends to produce clusters of similar size, but in real data one would normally expect the proportion of cheaters to be well below 50%. Secondly, k-means with $k = 2$ produces a hyperplane that divides the test-takers into two groups; so if the cheaters cannot be divided from the non-cheaters in this way, they will not be correctly identified. After finding the clusters, Man et al manually looked at person-fit and response-time variables to decide which cluster seemed to be the cheaters.

Future directions of research

For cluster analysis and the other ML methods described above, further research is needed if they are to become established. Countless other ML methods are available (Hastie et al 2009) and might prove useful. There is also a need to consider the broader issues.

Firstly, from the point of view of an ML scientist, the obvious primary task would be to predict whether a test-taker is cheating, rather than to explore distributions or model one dataset at a time. For this prediction task it is natural to use supervised rather than unsupervised ML. This would have the test-takers' responses, and possibly other data, as input, and a judgement of which test-takers were cheaters as the output. However, supervised ML usually needs high-quality data, in which it is known with high confidence who cheated and who did not. For cheating on tests, such data is often not available, and even when it is, the patterns cannot be assumed to generalise from one administration of a test to another.

Another problem with ML is that many methods are difficult or impossible to interpret. The approach of both Thomas (2016) and Man, Sinharay and Harring (2018) was essentially to put variables into a black box and see what happens. This often means that initial exploratory analysis is omitted, and the distributions of the variables or their sets of possible values are not considered. This approach is versatile and sometimes works well (Hastie et al 2009:590), and some ML methods involve thousands of parameters and complex interactions and so it is natural to treat them as a black box. But at least two of the algorithms used by Man et al (2019) (SVMs and random forests) can be relatively easy to interpret – the relations between the input and output variables can be described verbally, and understood without the need for technical knowledge. Greater consideration of the data,

the distributions, the algorithms, and their parameters, options, and variants might lead to better results.

For cluster analysis, one obvious avenue for future research is to use different clustering methods. One fast and well-established method is DBSCAN (Ester, Kriegel, Sander and Xu 1996). With DBSCAN, clusters are defined not by their centres but by their density. The basic idea is that if an object is within a certain distance of a certain number of objects, then all those objects are placed together in a cluster (the actual definition is slightly more complicated, to deal with objects inside the cluster and objects on the edge of the cluster). The distance threshold is called Eps and the minimum number of objects within that distance is called MinPts. These two parameters have to be specified but they are usually easier to choose than the number of clusters for k-means – if it is believed that a cluster ought to contain at least 10 objects, this can be achieved by simply setting MinPts to 10. DBSCAN is fast and flexible. Another possible advantage is that objects that are not close to any others are not put into a cluster; so if DBSCAN was used for cheating-detection, all clusters could be flagged as suspicious.

Most clustering methods do not identify a set of test-takers who may have cheated, so the three methods described in the previous section all used a second stage in which each cluster was classified as either cheaters or non-cheaters (the same was done with the other unsupervised methods in Man et al 2019). This second stage essentially converted the unsupervised algorithm into a supervised one. In Maynes and Thomas (2017) the second stage was automated, in Man et al (2019) it was manual, and in Wollack and Maynes (2017) all clusters were classified as possible cheaters. The two-stage procedure can be regarded as one overall statistical model or algorithm. One avenue for future research might be to look at the properties and implications of that overall model, and make a closer connection between the IRT model, similarity index, or other statistics in the first stage and the calculations in the second stage.

Conclusion

Language tests are more popular than ever, and the results often determine whether test-takers can get into university, enter a profession, or migrate internationally. The stakes are high, new ways of cheating are emerging, and researchers and testing organisations are coming up with new ways to combat them.

The first line of defence is test security (Wollack and Fremer (Eds) 2013a), but post-test screening and investigation are also indispensable. Statistical methods are well established and machine learning seems to hold promise. Unresolved issues include how to deal with the output of unsupervised methods and whether supervised methods might be preferable. More broadly, there is the question of whether to prefer statistical models,

moderately transparent machine learning methods, or a black box approach. Black box methods entail ethical issues – if a test-taker appeals against a judgement that they cheated, it would probably not be satisfactory for the testing organisation to be unable to explain how the judgement was made.

When evaluating these approaches and methods, a major and perennial problem is that it is often difficult or impossible for a researcher or testing organisation to find out for certain whether a test-taker cheated or not. Sometimes a test-taker confesses, or there are multiple pieces of evidence that they cheated, but usually there is no black-and-white proof. In ML terms, there is no good training data. Consequently, by far the most common way to evaluate these methods has been simulated data.

The lack of certainty about whether a test-taker cheated does not need to be a problem for testing organisations or programmes when applying these methods. As recommended by Cizek and Wollack (2017a), the focus can be on certifying or not certifying the test scores as trustworthy, based on the statistical evidence, rather than making a judgement about whether the test-taker followed the rules or cheated. If there is sufficient evidence that the score is invalid, then it is withheld. For example, Cambridge Assessment English has published its procedure for investigating cases of suspected cheating and states that this 'may lead to results being permanently withheld' (Cambridge Assessment English 2021).

Software may be a problem, since for many of the published methods there are no packages available and the only way is to write code. However, two packages for the R statistical programming language are available, for calculating person-fit and similarity indices (Tendeiro, Meijer and Niessen 2016, Zopluoglu 2018).

In their predictions for the future, Fremer and Wollack (2013) list technological developments as one of the key issues in the prevention and detection of cheating. They mention the increase in computer-based tests and the decreasing cost of biometric technologies for verifying test-takers' identities, and point out that some test venues are less sophisticated than others. These issues are especially important for language tests, which are often taken in a wide variety of locations around the world.

For further reading, books about cheating on tests include Cizek (1999), Wollack and Fremer (2013a), Kingston and Clark (Eds) (2014), and Cizek and Wollack (Eds) (2017b). Many articles and interviews are available from the Caveon Security Insights blog (Caveon 2019).

References

Albanese, M and Wollack, J (2013) *Cheating on tests: A threat to response validity*, paper presented at the 2nd annual Statistical Detection of Potential Test Fraud Conference, 19 October 2013.

American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME] and Joint Committee on Standards for Educational and Psychological Testing (1999) *Standards for Educational and Psychological Testing*, Washington, D.C.: American Educational Research Association.

American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME] and Joint Committee on Standards for Educational and Psychological Testing (2014) *Standards for Educational and Psychological Testing*, Washington, D.C.: American Educational Research Association.

Angoff, W H (1974) The development of statistical indices for detecting cheaters, *Journal of the American Statistical Association* 69 (345), 44–49.

BBC (2014) *Panorama – Immigration Undercover: The Student Visa Scandal*, available online: www.bbc.co.uk/programmes/b006t14n

BBC (2019) *US college admissions scandal*, available online: www.bbc.co.uk/news/topics/cdmk2l107p2t

Belov, D I and Wollack, J A (2018) *Detecting Groups of Test Takers Involved in Test Collusion as Unusually Large Cliques in a Graph (RR 18-01)*, available online: www.lsac.org/data-research/research/detecting-groups-test-takers-involved-test-collusion-unusually-large-cliques

Benjamini, Y and Hochberg, Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1), 289–300.

Benjamini, Y and Yekutieli, D (2001) The control of the false discovery rate in multiple testing under dependency, *Annals of Statistics* 29 (4), 1,165–1,188.

Bindmans (2016) *TOEIC Common Documents Bundle*, available online: www.bindmans.com/uploads/files/documents/TOEIC_Common_Documents.pdf

Bird, C (1927) The detection of cheating in objective examinations, *School and Society* 25 (635), 261–262.

Bird, C (1929) An improved method of detecting cheating in objective examinations, *The Journal of Educational Research* 19 (5), 341–348.

Bishop, N S, Huff, K, Mitchell, K, Rose-Bond, S, Stemmer, P, Trent, E R and Wollack, J (2012) *Testing and data integrity in the administration of statewide student assessment programs*, available online: archive.org/details/ERIC_ED605011/page/2/mode/2up

Bishop, S and Egan, K (2017) Detecting erasures and unusual gain scores, in Cizek, G J and Wollack, J A (Eds) *Handbook of Quantitative Methods for Detecting Cheating on Tests*, New York: Routledge, 193–213.

Boughton, K A, Smith, J and Ren, H (2017) Using response time data to detect compromised items and/or people, in Cizek, G J and Wollack, J A (Eds) *Handbook of Quantitative Methods for Detecting Cheating on Tests*, New York: Routledge, 177–190.

Bowers, M J, Wilson, R E and Hyde, R L (2011) *Georgia investigation into cheating in Atlanta Public Schools*, available online: archive.org/details/215260-georgia-investigation/page/n11

Cambridge Assessment English (2021) *Cambridge English Malpractice Procedure*, available online: www.cambridgeenglish.org/help/malpractice/

Caveon (2019) *Caveon security insights blog*, available online: caveon.com/news/caveon-security-insights/

Cizek, G J (1999) *Cheating on Tests*, Mahwah: Lawrence Erlbaum Associates.

Cizek, G J and Wollack, J A (2017a) Exploring cheating on tests, in Cizek, G J and Wollack, J A (Eds) *Handbook of Quantitative Methods for Detecting Cheating on Tests*, New York: Routledge, 3–19.

Cizek, G J and Wollack, J A (Eds) (2017b) *Handbook of Quantitative Methods for Detecting Cheating on Tests*, New York: Routledge.

Cohen, A S and Wollack, J A (2006) Test administration, security, scoring, and reporting, in Brennan, R L (Ed) *Educational Measurement* (Fourth edition), Westport: Praeger Publishers, 355–386.

Contreras, P and Murtagh, F (2016) Hierarchical clustering, in Hennig, C, Meila, M, Murtagh, F and Rocci, R (Eds) *Handbook of Cluster Analysis*, Florida: CRC Press, 103–123.

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

Dudley, R, Stecklow, S, Harney, A and Liu, I J (2016) *As SAT was hit by security breaches, College Board went ahead with tests that had leaked*, available online: www.reuters.com/investigates/special-report/college-sat-one/

Dunn, O J (1961) Multiple comparisons among means, *Journal of the American Statistical Association* 56 (293), 52–64.

Ester, M, Kriegel, H-P, Sander, J and Xu, X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*.

Färber, I, Günnemann, S, Kriegel, H-P, Kröger, P, Müller, E, Schubert, E, Seidl, T and Zimek, A (2010) *On using class-labels in evaluation of clusterings*, available online: eecs.oregonstate.edu/research/multiclust/Evaluation-4.pdf

Foster, D (2013) Security issues in technology-based testing, in Wollack, J A and Fremer, J J (Eds) *Handbook of Test Security*, New York: Routledge, 9–84.

Foster, D, Mattoon, N and Shearer, R (2009) *Using multiple online security measures to deliver secure exams to test takers. A white paper*, available online: www.caveon.com/wp-content/uploads/2014/03/KRY_WhitePaper_OLPFinal.pdf

Fremer, J J and Wollack, J A (2013) Conclusion: The future of test security, in Wollack, J A and Fremer, J J (Eds) *Handbook of Test Security*, New York: Routledge, 343–350.

French, P (2016) *Report on Forensic Speaker Comparison Tests Undertaken by ETS*, available online: www.bindmans.com/uploads/files/documents/TOEIC_Common_Documents.pdf#page=61

Gentleman, A (2019) Home Office faces legal action over English test cheating claims, *The Guardian*, available online: www.theguardian.com/uk-news/2019/apr/26/home-office-faces-legal-action-over-english-test-cheating-claims

Geranpayeh, A (2013) Detecting plagiarism and cheating, in Kunnan, A (Ed) *The Companion to Language Assessment*, New Jersey: John Wiley & Sons Inc, 980–993.

Hartigan, J A and Wong, M A (1979) Algorithm AS 136: A K-Means Clustering Algorithm, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28 (1), 100–108.

Hastie, T, Tibshirani, R and Friedman, J (2009) *The Elements of Statistical Learning* (Second edition), New York: Springer Science+Business Media.

Hazell, W (2019) *Exclusive: Exam cheating commission wants to 'future proof' the system*, available online: www.tes.com/magazine/archive/exclusive-exam-cheating-commission-wants-future-proof-system

Hennig, C and Meila, M (2016) Cluster analysis: An overview, in Hennig, C, Meila, M, Murtagh, F and Rocci, R (Eds) *Handbook of Cluster Analysis*, Florida: CRC Press, 1–20.
iTEP (2018) *Test Security*, available online: www.itepexam.com/admissions/test-security/
Joint Council for Qualifications (2018) *Media Release: JCQ announces an Independent Commission into Malpractice*, available online: www.jcq.org.uk/jcq-announces-an-independent-commission-into-malpractice/
Karabatsos, G (2003) Comparing the aberrant response detection performance of thirty-six person-fit statistics, *Applied Measurement in Education* 16 (4), 277–298.
Kim, D, Woo, A and Dickison, P (2017) Identifying and investigating aberrant responses using psychometrics-based and machine learning-based approaches, in Cizek, G J and Wollack, J A (Eds) *Handbook of Quantitative Methods for Detecting Cheating on Tests*, New York: Routledge, 70–97.
Kingston, N M (2013) Educational testing case studies, in Wollack, J A and Fremer, J J (Eds) *Handbook of Test Security*, New York: Routledge, 299–312.
Kingston, N M and Clark, A K (Eds) (2014) *Test Fraud: Statistical Detection and Methodology*, New York: Routledge.
Kriegel, H-P, Schubert, E and Zimek, A (2017) The (black) art of runtime evaluation: Are we comparing algorithms or implementations?, *Knowledge and Information Systems* 52 (2), 341–378.
Main, E (2016) *Home Office under fire after student visa fraud case*, available online: www.bbc.co.uk/news/uk-38236852
Man, K, Harring, J R and Sinharay, S (2019) Use of data mining methods to detect test fraud, *Journal of Educational Measurement* 56 (2), 251–279.
Man, K, Sinharay, S and Harring, J R (2018) *Use of data mining methods to detect test fraud*, paper presented at National Council on Measurement in Education, New York, 12–16 April 2018.
Maslen, G (1996) *Cheats with pagers and cordless radio cribs*, available online: www.tes.com/news/cheats-pagers-and-cordless-radio-cribs
Maynes, D (2014) Detection of non-independent test taking by similarity analysis, in Kingston, N and Clark, A (Eds) *Test Fraud: Statistical Detection and Methodology*, New York: Routledge, 53–82.
Maynes, D D (2017) Detecting potential collusion among individual examinees using similarity analysis, in Cizek, G J and Wollack, J A (Eds) *Handbook of Quantitative Methods for Detecting Cheating on Tests*, New York: Routledge, 47–69.
Maynes, D (2018) *Celebrating 15 Years: An Interview with Dennis Maynes*, available online: blog.caveon.com/celebrating-15-years-dennis-maynes
Maynes, D D and Thomas, S L (2017) *A method for estimating the latent source used by answer copiers*, paper presented at National Council on Measurement in Education, San Antonio, 26–30 April 2017.
Mirkin, B (2016) Quadratic error and k-means, in Hennig, C, Meila, M, Murtagh, F and Rocci, R (Eds) *Handbook of Cluster Analysis*, Florida: CRC Press, 33–54.
National Audit Office (2019) *Investigation into the response to cheating in English language tests*, available online: www.nao.org.uk/reports/investigation-into-the-response-to-cheating-in-english-language-tests/
National Council on Measurement in Education (2019) *National Council on Measurement in Education Position Statement on Test Security for Large*

Scale Educational, Credentialing, and Workplace Testing, available online: www.ncme.org/resources-publications/position-statements/measurement-practices

Noah, H J and Eckstein, M A (2001) *Fraud and Education: The Worm in the Apple*, Lanham: Rowan & Littlefield Publishers.

Qian, H, Staniewska, D, Reckase, M and Woo, A (2016) Using response time to detect item preknowledge in computer-based licensure examinations, *Educational Measurement: Issues and Practice* 35 (1), 38–47.

Saupe, J L (1960) An empirical model for the corroboration of suspected cheating on multiple-choice tests, *Educational and Psychological Measurement* 20 (3), 475–489.

Scott, M, Cooper, C and Maynes, D (2015) *Analysis of flawed answer keys to detect braindump usage*, paper presented at Conference on Test Security 2015, 4–6 November 2015.

Shi, G R (1993) Multivariate data analysis in palaeoecology and palaeobiogeography—a review, *Palaeogeography, Palaeoclimatology, Palaeoecology* 105 (3), 199–234.

Skorupski, W P and Wainer, H (2017) The case for Bayesian methods when investigating test fraud, in Cizek, G J and Wollack, J A (Eds) *Handbook of Quantitative Methods for Detecting Cheating on Tests*, New York: Routledge, 346–357.

Tendeiro, J N, Meijer, R R and Niessen, S M (2016) PerFit: An R package for person-fit analysis in IRT, *Journal of Statistical Software* 74 (5), 1–27.

Thomas, S (2016) *So Happy Together? Combining Rasch and Item Response Theory Model Estimates with Support Vector Machines to Detect Test Fraud*, available online: libraetd.lib.virginia.edu/public_view/4m90dv85v

United States Department of Justice (2019a) *Arrests Made in Nationwide College Admissions Scam: Alleged Exam Cheating & Athletic Recruitment Scheme*, available online: www.justice.gov/usao-ma/pr/arrests-made-nationwide-college-admissions-scam-alleged-exam-cheating-athletic

United States Department of Justice (2019b) *Investigations of College Admissions and Testing Bribery Scheme*, available online: www.justice.gov/usao-ma/investigations-college-admissions-and-testing-bribery-scheme

Van der Linden, W J (2007) A hierarchical framework for modeling speed and accuracy on test items, *Psychometrika* 72 (3), 287–308.

Wasserman, L (2004) *All of Statistics: A Concise Course in Statistical Inference*, New York: Springer-Verlag.

Watkins, D and Biggs, J (1996) *The Chinese Learner: Cultural, Psychological, and Contextual Influences*, Hong Kong: Comparative Education Research Centre, Faculty of Education, University of Hong Kong.

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

Weir, C J and Shaw, S (2005) Establishing the validity of Cambridge ESOL Writing tests: Towards the implementation of a socio-cognitive model for test validation, *Research Notes* 21, 10–14.

Wollack, J A (1997) A nominal response model approach for detecting answer copying, *Applied Psychological Measurement* 21 (4), 307–320.

Wollack, J A (2018) *Ask an Expert: James A Wollack*, available online: thelockbox.readz.com/ask-an-expert-jim-wollack

Wollack, J A and Fremer, J J (Eds) (2013a) *Handbook of Test Security*, New York: Routledge.

Wollack, J A and Fremer, J J (2013b) Introduction: The test security threat, in Wollack, J A and Fremer, J J (Eds) *Handbook of Test Security*, New York: Routledge, 1–13.

Wollack, J A and Maynes, D D (2017) Detection of test collusion using cluster analysis, in Cizek, G J and Wollack, J A (Eds) *Handbook of Quantitative Methods for Detecting Cheating on Tests*, New York: Routledge, 124–150.

Wollack, J A and Schoenig, R W (2018) Cheating, in Frey, B B (Ed) *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*, Thousand Oaks: Sage, 260–265.

Wollack, J A, Cohen, A S and Serlin, R C (2001) Defining error rates and power for detecting answer copying, *Applied Psychological Measurement* 25 (4), 385–404.

Xi, X (2010) How do we go about investigating test fairness?, *Language Testing* 27 (2), 147–170.

Zopluoglu, C (2017) Similarity, answer copying, and aberrance: Understanding the status quo, in Cizek, G J and Wollack, J A (Eds) *Handbook of Quantitative Methods for Detecting Cheating on Tests*, New York: Routledge, 25–46.

Zopluoglu, C (2018) *CopyDetect: Computing Response Similarity Indices for Multiple-choice Tests*, Version 1.3, available online: cran.r-project.org/src/contrib/Archive/CopyDetect/

Epilogue

Guoxing Yu
University of Bristol, UK

Jing Xu
Cambridge University Press & Assessment, UK

In the Introduction to this volume, we briefly reviewed the motives, benefits and challenges of using technology in language assessment from historical perspectives. The studies included in this edited volume are broadly within two main areas of technology use in language assessment, i.e., using technology to validate language assessment tasks, and using technology to enhance language assessment task development, delivery, and examiner and test-taker experience. In this Epilogue, we will present a broad sketch of what we have learned from these studies and point to future directions of technology use in language assessment.

What have we learned from the studies reported in this edited volume?

Part 1 includes eight chapters. Six of them reported empirical studies that used either eye tracking or keystroke logging as the main data collection tool to investigate test-takers' cognitive processes in task completion. Another empirical study (Chapter 7), alongside collecting eye-movement data, explored test-takers' attention to their assessment feedback. One chapter (Chapter 8) presented a literature review of two data collection tools, eye tracking and electroencephalography (EEG), both of which can be used to tap into test-takers' cognitive processes during task completion. As shown in these chapters, there is an increasing interest in using eye tracking as a data collection tool to investigate test-taking processes. This strand of validation and research studies in the field of language assessment has been informed methodologically by research in cognitive psychology, psycholinguistics, and language processing and production. What is evident from all these studies, however, is the importance of incorporating and triangulating data from other sources to better understand test-takers' cognitive processes in task completion and their attention and engagement with assessment feedback reports. **Part 2** includes five chapters. Four of them were empirical studies that used a range of technology (e.g., video-conferencing, automated speech recognition, machine learning) to develop assessment tasks and

automated marking, and diagnostic and feedback systems. Although they were all small-scale and explorative in their research scope and approach to utilising technology, these empirical studies demonstrated the potential of technology to expedite innovations and increase efficiency and accessibility of assessment tools for both high-stakes and low-stakes purposes. The final chapter introduced statistical and computational methods to detect cheating in objectively marked tests. It is a good example of how technology can be used to ensure the integrity of assessment systems. The 11 empirical studies and the two literature reviews included in this edited volume offer some unique and timely insights into what technology can currently do, in different educational contexts and for different purposes, to facilitate certain advancements in language test design, delivery, and validation. However, such advancements are limited and incremental in their overall scope and reach, partly because of the limitations of technology employed in these studies and partly because of different views people hold of the role of technology in language assessment (i.e., whether technology is an integral part of the construct of language assessment or technology plays only a facilitation role to improve the efficiency of assessment).

Future directions of technology use in language assessment research and practice

Given the rapid change of technology, it is difficult if not impossible to predict what technology can do and how it would affect language assessment research and practice. At the initial stage of this book project, for example, artificial intelligence (AI) was hardly a phenomenon, but AI is now playing an increasingly important, and arguably disruptive, role in language assessment research and practice. Like any other disruptive technology, AI, though still emerging, is already making transformative changes to the technological landscape and methods in areas such as automated marking of written and spoken language production, automated and personalised feedback on test performance, automated generation of assessment tasks in different modes (text, still images, video, and audio), automated calibration of test questions, and automated detection and prevention of cheating during remote invigilation and/or based on the analysis of test performance data. AI will continue to do so, but perhaps at a much more rapid speed. In this line of future directions, technology is being used mainly to improve efficiency and user experience of language assessments.

In the second line of future directions, we argue that technology use has become an integral part of the construct of language assessment, rather than playing a mediating or facilitating role. Technology is changing the way we communicate and learn a language. It is widely acknowledged that technology use is essential in our increasingly multimodal communication,

and that language is part and parcel of visual, auditory, and spatial patterns of meaning-making in communication. Technology makes it possible to design and deliver multimodal assessment tasks that can better reflect human-to-human and human–AI communications in real life, thus making test-takers' experience more intuitive, interactive, and immersive. Therefore we call for, as Yu and Zhang (2017) and Taylor and Banerjee (2023) did, reconceptualising the construct of language assessment in a digital age. The expansion and reconceptualisation of the construct of language assessment to include technology use can open an array of opportunities for designing innovative technology-integrated assessment tasks.

In the third line of future directions, we envisage technology being used to address issues related to EDI (equality, diversity, and inclusion). Personalised, digitally presented assessment tasks would offer opportunities for test-takers with special requirements to be accommodated with more suitable and accessible assessment tasks so that they can perform to the best of their abilities. In this line of future directions, technology plays double roles: mediating/facilitating task presentation and completion, and being part of the construct of language assessment because test-takers with special requirements often rely on assistive technology for their communication. As O'Sullivan (2023:509) argued, we must leverage 'the potential of technology to help us deliver truly personalised, communication-oriented tests to as broad an audience as possible'.

The fourth line of future directions serves as our caution, given that this edited volume is largely technology-driven without due consideration of the sociological aspects of educational assessment. Despite our enthusiasm in using technology in language assessment research and practice, we are keenly aware of and must acknowledge the limitations of technology itself as well as the 'legitimatory' roles (Broadfoot 1996) and social functions of educational assessment in mediating the relationship between education and society. In Broadfoot's (1996) view, which we heartily embrace, educational assessment is the attestation of learners' *competence*; it regulates *competition* for opportunities and resources; determines the *content* of teaching and learning; and serves as a rational means of *control* of educational quality, social mobility, and reform, among other purposes. Therefore, we argue that any future directions of technology use in language assessment research and practice should be closely scrutinised from such sociological approaches to educational assessment.

References

Broadfoot, P (1996) *Education, Assessment and Society*, Maidenhead: Open University Press.

O'Sullivan, B (2023) Reflections on the application and validation of technology in language testing, *Language Assessment Quarterly* 20 (4–5), 501–511.

Taylor, L and Banerjee, J (2023) Accommodations in language testing and assessment: Safeguarding equity, access, and inclusion, *Language Testing* 40 (4), 847–855.
Yu, G and Zhang, J (2017) Computer-based English language testing in China: Present and future, *Language Assessment Quarterly* 14 (2), 177–188.