Trends in Language Assessment Research and Practice

Trends in Language Assessment Research and Practice:

The View from the Middle East and the Pacific Rim

Edited by

Vahid Aryadoust and Janna Fox

Cambridge
Scholars
Publishing

Trends in Language Assessment Research and Practice:
The View from the Middle East and the Pacific Rim

Edited by Vahid Aryadoust and Janna Fox

This book first published 2016

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-8261-5
ISBN (13): 978-1-4438-8261-3

TABLE OF CONTENTS

Part 2: Trends in Understanding Individual Variation in Assessment Outcomes: Test Takers and Raters

Part 3: Trends in Diagnostic Assessment: Its Role in Supporting Learning and Academic Success

FOREWORD

ANDY CURTIS
GRADUATE SCHOOL OF EDUCATION, ANAHEIM UNIVERSITY, CALIFORNIA, USA
PRESIDENT (2015-2016), TESOL INTERNATIONAL ASSOCIATION
LIYING CHENG
PROFESSOR AND DIRECTOR OF ASSESSMENT AND EVALUATION GROUP (AEG)
FACULTY OF EDUCATION, QUEEN'S UNIVERSITY, ONTARIO, CANADA

Trends in language assessment research and practice: The view from the Middle East and the Pacific Rim is not only a book about recent and current developments in language assessment, it is a book about the future of the field and the essential role that context plays in test development and validation. That said, predictions about the future are notoriously prone to being unreliable, so before such a bold claim is made, let us take a very brief look at our past.

At the start of our most recent new millennium, in January 2000, Lyle Bachman's article 'Modern language testing at the turn of the century: assuring that what we count counts', appeared in *Language Testing* (Volume 17). The article is a 40-plus-page comprehensive review of language testing in the 1980s and 1990s, drawing on more than 200 published works. This personal, retrospective paper focuses on a wide range of areas that developed mainly in the 1990s, including various characteristics. For example, Bachman noted that: "Kunnan (1998a) lists over 20 studies investigating test-taker characteristics, such as academic background, native language, culture, gender, and field dependence. Other characteristics that have been studied include occupation (Hill, 1993), aptitude (Sasaki, 1996; Sparks et al., 1998), background knowledge (Clapham, 1993; 1996) and personality characteristics (Berry, 1993)" (Bachman, 2000, p.11). However, what is conspicuous by its absence, in

these lists and in the article itself, is any reference to the geographical context as an essential aspect of the language testing taking place, and yet, *where* the language testing is happening is potentially as influential a factor in the test results as any of those listed above. Language testing and assessment, like language teaching, do not take place in a vacuum. They all happen somewhere – and that somewhere matters.

The long-standing importance of context – a main feature of this current book – is one of the characteristics that most clearly distinguishes language testing from language teaching, as the latter has a history of up to 5,000 years (Germain, 1993), and a well-documented history of 25 centuries, from 500 BC to 1969 (Kelly, 1969) English language teaching has a history going back well over 600 years, to the 1400s (Howatt & Widdowson, 2004). That is in sharp contrast to the much shorter history of language testing, which, according to Antony Kunnan's editorial for the inaugural issue of *Language Assessment Quarterly*, dates back less than a century: "The recorded study of modern language testing arguably began in the 1930s and 1940s" (2004, p.2).

Further, one of the key differences between language testing and language assessment may be the awareness of the importance of context. According to Glenn Fulcher and Fred Davidson (2007): "The main difference between classroom assessment and large-scale educational assessment is the context of the classroom" (p.24), which we would re-phrase as: The main difference between large-scale educational testing and assessment is the context of the classroom – and not only the classroom, but also the country or region, with its distinctive political, social and pedagogical variability. As elegantly stated by Caroline Clapham (2000, p.147), "the relationship between language testing and the other sub-disciplines of applied linguistics" is indeed "the relationship between testing and assessment".

This brings us back to our claim that *Trends in language assessment research and practice: The view from the Middle East and the Pacific Rim* is not only a book about recent and current developments in language assessment, and a book about the future of the field, but it is also unique in the ways in which it focuses on context. The two editors, Vahid Aryadoust and Janna Fox, one based in Singapore and one in Canada, have brought together more than 40 contributors from all over the world, representing a truly international collaboration, with perspectives on – and, in most cases, data from – Abu Dhabi, Australia, Canada, Egypt, Germany, Hong Kong,

Iran, Japan, South Korea, Mainland China, New Zealand, Singapore, the USA, Ukraine, and elsewhere. There are 26 chapters in this extensive volume. Below we discuss a few of the contributions to provide a sense of the range and depth of scholarship included here.

Aryadoust and Fox note in their Introduction to this collection that, "modern language assessment is intensely interdisciplinary" (p.1) and advancing "from multiple directions" (p.2). This relates to another one of the ways in which this book will move our field forward in important new directions, by bringing techniques and approaches not previously associated with language assessment into the repertoire. See, for example Fox and Hirotani, Chapter Five, and their use of Functional Magnetic Resonance Imaging (fMRI) in language testing research. fMRI is a relatively recent non-invasive technology used to measure and to map brain activity. The introductory comments of Aryadoust and Fox, on the interdisciplinary and multidirectional nature of modern language assessment, and contributions such as Chapter 5, show where we are now, and where we are headed, in relation to Bachman's reflections, more than 25 years ago, on the previous two decades: "In the past twenty years [1980-2000] language testing research and practice have witnessed the refinement of a rich variety of approaches and tools for research and development, along with a broadening of philosophical perspectives and the kinds of research questions that are being investigated" (Bachman, 2000, p.1). However, in addition to the ways in which the approaches, research and development tools, perspectives, and questions have changed, a more global perspective – geographically, disciplinarily, philosophically, and methodologically – is now evident.

As the co-editors of *Trends in language assessment research and practice* point out research on language assessment from the Pacific Rim and the Middle East have, so far, "been underrepresented in … scholarly work" (p.2), making this book "a unique overview of contemporary research in language assessment in these two regions" (p.2). We agree with the co-editors' introductory statement that: "The trends identified in this volume and the research reported here suggest that researchers across the Middle East and the Pacific Rim are playing – and will continue to play – an important role in advancing the quality, utility, and fairness of language testing and assessment practices" (p.10).

In terms of new and emerging trends, this book presents a number of promising developments for the future. For example, in Chapter 2,

Aryadoust concludes that: “Due to its significantly high accuracy across learning and validation sets, symbolic regression can be used as an emergent technique in language assessment … for depicting the cognitive mechanisms of writers” (p.44). Moving beyond large-scale, standardized, commercial language testing to classroom assessment brings individual language teachers and learners into focus, who may benefit from the research reported in this book. For example, Kinnie Kin Yee Chan and Trevor Bond argue that, “the Lexile Analyzer could be used to help lighten Hong Kong language teachers’ grading load and release some of their considerable stress in essay scoring processes” (p.68).

As noted above, Fox and Hirotani not only highlight the “exciting new tools available to SLA and LT researchers” (Ch. 5, p.92) such as fMRI and Event Related Potentials (ERP), they reiterate the point that, “Interdisciplinary collaboration will greatly benefit language testing research in future” (Ch. 5, p.109), and they remind us that “second language learning is … contextual and cultural” (Ch. 5, p.92). This is one of the most prominent recurring themes in this book, whether it is Alan Urmston and Michelle Raquel referring to Hong Kong’s “unique blend of Eastern and Western influences” (p.232), wherein learners have, “generally low levels of English proficiency and have negative attitudes towards English learning” (p.232), or Australia, New Zealand and Canada, where a “Debate is currently raging in many educational jurisdictions over underprepared, at-risk students in university” (Ch. 12, Fox, von Randow and Volkov, p.266).

In one of the most effective illustrations of the importance of context in language assessment, Roger Nunn and John Langille, in Chapter 14, report on their work at the Petroleum Institute in Abu Dhabi in the United Arab Emirates (UAE). In their chapter, they “highlight the contrast between the traditional approach sometimes associated with the Middle East and attempts to operationalize a more holistic and global view of competence … in our local context” (Ch. 14, p.302). Nunn and Langille also “consider the relationship between local communities such as our own and more global communities of practice” (Ch. 14, p.303), and explain why such work is necessary, because “our local students will increasingly need to interact with scholars from other contexts both within our own multicultural local community and beyond” (Ch. 14, p.303). With parts of this chapter subtitled “The Local Context: The Battle of the Pedagogies – Ongoing” and “Adaptation to a Local Context”, as well as “A Local Task Description” and “Local in Contrast to Global Competence”, Nunn and

Langille conclude by explaining that: "Our characterization is globally expressed, but locally illustrated and is an approach to holistic in-house assessment" (Ch. 14, p.314).

Another important global shift has been the move beyond English and other European languages, such as French and German, to the assessment of other languages, elsewhere. An example of this is presented in Chapter 15, in which Wu Siew Mei and Susan Tan, in Singapore, point out that "Because of the increasing global influence of China, the Chinese language, a part of China's soft power, has gained unprecedented popularity around the world", as seen in the "global 'Hanyu Re' (Chinese Fever), or craze for learning Chinese as a Second Language (CSL)" (p.318). Another good example of the importance of contextual factors is seen in Chapter 19, in which Atta Gebril, Deena Boraie and Elizabeth Arrigoni report on 'Assessment literacy training for English language educators in Egypt' (pp.416-437). Gebril, Boraie and Arrigoni provide "useful guidelines for teacher trainers working in test-driven instructional contexts similar to Egypt, as well as low-resource contexts in which other models of training are less feasible" (p.416).

In terms of how such research can help classroom teachers, which is another one of the recurring themes in this book, Gebril, Boraie and Arrigoni explain that one of their goals was that: "Given this context, an assessment literacy training program was proposed for English language educators in Egypt as an empowerment strategy towards education reform through developing the assessment knowledge, skills, and attitudes of language educators in Egypt" (p.419). Another good example of helping teachers in resource poor environments is presented in Chapter 21, by Cynthia Wiseman, on the washback effects of the *gao kao* – "a battery of standardized summative assessments administered in primary, middle, and secondary schools in China" (p.464) – on communicative language teaching in English language classrooms across China.

The scale of English language testing in China is bigger than anything anywhere else, with a single examination, the College English Test (CET), taken by more than 18 million students annually in China (Yu & Jin, 2014, see also Cheng and Curtis, 2010). Given such mind-boggling numbers, it is no surprise that this situation "exerts tremendous impact on university/college teaching and learning of English in China and affects huge numbers of stakeholders" (p.468). In the language testing of a single country, on such an enormous scale – unprecedented in human history – it

can be hard to know what to do to help. But Wiseman's chapter, "provides insight into the pedagogical approach and practices commonly implemented in rural EFL classrooms in China to inform teacher training, and perhaps more importantly, it gives voice to those 'on the front line' – the educators teaching in the public school, day in and day out" (pp.477-478). Wiseman concludes that: "In spite of the challenges, or perhaps because of them, the work of rural teacher training programs is vital" (p.481).

The last chapter of the book (Ch. 26), by Trevor Bond, is, like Bachman's 2000 paper in *LT*, a more personal reflection, in which he discusses some powerful influences not mentioned elsewhere in the book, such as the fact that "Testing is Big Business" (pp.576-577). Bond also reminds us of certain 'universal truths', such as the fact that: "Language test scores have consequences for the candidates; sometimes, life-changing consequences in high-stakes testing for, say, university entrance/exit or professional accreditation for those with English as a second or foreign language" (p.583). Bond concludes with a reminder of what happens when we get it wrong: "Although the impact of wrong placement test allocation might be minor for some, it is likely that the consequences of test scores unaided by such guided professional reflection could be quite calamitous" (p.590).

Trends in language assessment research and practice: The view from the Middle East and the Pacific Rim not only brings the most recent work on language assessment in those parts of the world to a wider audience, but also makes many new and significant contributions to the role that context plays in test development and validation, which we believe have the potential to help language teachers and learners all over the world.

—Andy Curtis and Liying Cheng
Ontario, Canada, October 2015

References

Bachman, L. F. (2000). Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing, 17(1),* 1-42.

Cheng, L., & Curtis, A. (Eds.). (2010). *English language assessment and the Chinese learner.* New York, NY: Routledge.

Clapham, C. (2000). Assessment and testing. *Annual Review of Applied Linguistics*, 20, 147-161.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book.* New York, NY: Routledge

Germain, C. (1993). *Evolution de l'Enseignement des Langues: 5000 Ans d'Histoire [Evolution of the teaching of languages: 5000 years of history]*. Paris, France: Clé International.

Howatt, A. P. R., & Widdowson, H. G. (2004). *A history of English language teaching.* Oxford, UK: Oxford University Press.

Kelly, L.G. (1969*). 25 Centuries of Language Teaching: An inquiry into the science, art, and development of language teaching methodology, 500 B.C.-1969.* Rowley, MA: Newbury House.

Kunnan, A. (2004). Regarding language assessment. *Language Assessment Quarterly, 1(1),* 1-4.

Yu, G. & Yan, J. (2014). *Editorial: English language assessment in China: policies, practices and impacts. Assessment in Education: Principles, Policy & Practice, 21(3),* 245–250.

FROM THE EDITORS

VAHID ARYADOUST
NATIONAL UNIVERSITY OF SINGAPORE, SINGAPORE
AND JANNA FOX
CARLETON UNIVERSITY, CANADA

Our interest in putting together the present volume grew out of a burgeoning stream of research into language assessment in the Middle East and the Pacific Rim. As the focus on education and the role of English language teaching continues to intensify across these regions at an unprecedented rate, assessing communication skills becomes an increasingly significant field. Some of the major universities in these regions have had a long history in teaching and assessing English and other languages, and researchers, practitioners, and scholars alike have attempted to develop innovative assessment approaches and techniques to address the pressing needs of language test developers and test takers. At the same time, multiple annual conferences, such as *Pacific Rim Objective Measurement Symposium* (PROMS) and *the Asian Association for Language Assessment* (AALA) conference, have been launched to bring scholars together and keep them updated about the latest developments in language and educational assessment in these regions.

Despite the prodigious developments in the field of language assessment in the Middle East and the Pacific Rim, we feel that research and practice in these regions have been underrepresented in mainstream literature and the present volume is an attempt to address this gap. To create this volume, we have tapped into the knowledge of some language and educational assessment experts whose diversity of perspectives and experience has enriched the focus and scope of language and educational assessment in general, and the present book in particular. We sincerely hope that the attempt of the contributors to promote and propagate new approaches and techniques in language assessment will help readers gain deeper insight into extant topics in language assessment, which are currently pursued in these regions, as well as new directions for future research.

We would like to thank all of the authors in the present volume for their stimulating contributions and their cooperation during the review process. We also wish to acknowledge that every chapter in this volume has been subjected to blind peer review in addition to review by the editors. As such, we wish to extend our thanks to the many language assessment colleagues, including the contributors, who helped us during the review process, and in particular, Elif Toprak, Hong Wang, Lin Chen, and Angel Arias. Finally, we wish to extend a special note of thanks to Ariele Pathi, our editorial assistant, for her conscientious efforts in insuring that the volume met the highest professional standards. We also extend our thanks to Alisa Zavialova and Kenson Tan, who provided extraordinary assistance in the final pre-publication stage.

ABBREVIATIONS AND ACRONYMS

Akaike Information Criterion (AIC)
American Council on the Teaching of Foreign Languages (ACTFL)
Artificial Neural Network (ANN)
Asian Association for Language Assessment (AALA)
Audio Lingual Method (ALM)
Autism-spectrum Quotient (AQ)
Automated Essay Scoring (AES)
Banked Cloze (BCLZ)
Bayesian Information Criterion (BIC
British Academic Written English (BAWE)
Business Chinese Test (BCT)
Canadian Academic English Language (CAEL)
Canadian Language Benchmarks (CLB)
Canadian Test of English for Scholars and Trainees (CanTEST)
Centre for English and Additional Languages (CEAL)
Children's Chinese Competency Certification (CCCC)
Chinese as a Second Language (CSL)
Chinese Proficiency Test or Hanyu Shuiping Kaoshi (HSK)
City University of Hong Kong (CityU)
Classical Test Theory (CTT)
Classification and regression trees (CART)
Cog = Cognitive
Cognitive Rater Types (CRTs)
College English Test (CET)
COM = Comprehending
Common English Proficiency Assessment Scheme (CEPAS)
Common European Frame of Reference (CEFR)
Communicative language learning (CLL)
Communicative Language Teaching (CLT)
Comparative fit index (CFI)
Confirmatory factor analysis (CFA)
Corpus of Contemporary American English (COCA)
D = Disturbance (in SEM)
Degrees of freedom (*df*)
Diagnostic English Language Assessment (DELA)

Diagnostic English Language Needs Assessment (DELNA)
Diagnostic English Language Tracking Assessment (DELTA)
Differential rater functioning (DRF)
E = Measurement error with the observed variable (in SEM)
Empathy Quotient (EQ)
English as a Foreign Language (EFL)
English as a Second Language (ESL)
English as the medium of instruction (EMI)
English for Academic Purposes (EAP)
English for Speakers of Other Languages (ESOL)
English for Specific Purposes (ESP)
English Language Learner (ELL)
English Language Proficiency Assessment (ELPA
English Language Teaching (ELT)
EVA = Evaluating
EVA = evaluating strategies
Event related potentials (ERP)
Exploratory factor analysis (EFA)
Functional magnetic resonance imaging (fMRI)
GEN = general progression strategies
General English Proficiency Test (GEPT)
General Test of English Language Proficiency (G-TELP)
Graduating Students' Language Proficiency Assessment (GSLPA)
GRAM = Grammatical
Hanyu Shuiping Kaoshi or Chinese Proficiency Test (HSK)
HK Polytechnic University (PolyU)
Hong Kong Diploma of Secondary Education (HKDSE)
Hong Kong Examinations and Assessment Authority (HKEAA)
Hong Kong Institute of Education (HKIEd)
Hong Kong Special Administrative Region (HKSAR)
IDE = identifying important information strategies
INF = inference-making strategies
Inference Use Argument (IUA)
Information Technology (IT)
INTE = integrating strategies
International English Language Testing System (IELTS)
International Language Testing Association (ILTA)
International Objective Measurement Workshops (IOMW)
Item Response Theory (IRT)
Japanese Ministry of Education, Sports, Science and Technology (MEXT)
Japan Association of Language Teachers (JALT)

Korean Scholastic Aptitude Test (KSAT)
Latent semantic analysis (LSA)
LEX-GR = lexico-grammatical reading ability
Lingnan University (LU)
Local item dependence (LID)
LR = long-range constraints items
Magnetic resonance imaging (MRI)
Many-facet Rasch measurement (MFRM)
MASUS = Measuring the Academic Skills of University Students
Maximum likelihood (ML)
Mean square (MNSQ)
MEM = Memory
Met = Metacognitive
Metacognitive and Cognitive Strategy Use Questionnaire (MCSUQ)
Minzu Hanyu Shuiping Kaoshi or Test of Chinese Proficiency for Minorities (MHK)
MON = Monitoring
MON = monitoring strategies
National Assessment of Educational Progress (NAEP)
National English Ability Test (NEAT)
National Matriculation English Test (NMET)
Native English Teacher (NET)
Native-English-speaker raters (NESs)
Necessary information (NI)
Neural Network (NN)
New Secondary School (NSS)
Non-Normed Fit Index (NNFI; also known as Tucker Lewis Index)
Normed chi-square (NC)
Normed Fit Index (NFI)
Objective Communicative Speaking Test (OCST)
Occupational English Test (OEC)
Onscreen Marking (OSM)
Operational rater types (ORTs)
Other Learning Experience (OLE)
Pacific Rim Objective Measurement Symposium (PROMS)
Pearson Test of English Academic (PTE)
People's Education Press (PEP)
PLA = planning strategies
PLAN = Planning
Post-Entry Diagnostic Assessment (PEDA)
Post-entry language assessment (PELA)

Posterior Superior Temporal gyrus (pSTG)
Practical English Level Test (PELT)
Primary medium of instruction (MOI)
Rasch Model Analysis (RSM)
Rasch testlet model (RTM)
Rasch-Andrich rating scale model (RSM)
Reading in-Depth (RID)
Receiver operating characteristic (ROC)
REM = reading for explicit meaning items
RET = Retrieval
RIM = reading for implicit meaning items
Root mean square error (RMSE)
Root Mean Square Error of Approximation (RMSEA)
S = State
Scholastic Assessment Test (SAT)
School-Based Assessment (SBA)
Second Language (L2)
Second Language Acquisition (SLA)
Self-Assessment for Engineering (SAFE)
Senior secondary (SS)
Skimming and Scanning (SKSN)
SKM = skimming items
SKN = scanning items
Special Educational Needs (SEN)
SR = short-range constraints items
Standardized (Zstd)
State Education Development Commission (SEDC)
Steering Committee for the Test of Proficiency-Huayu (SC-TOP)
STR_U = strategy use
Structural equation modelling (SEM)
Systemizing Quotient (SQ)
T = Trait
Teachers of English to Speakers of Other Languages (TESOL)
Territory-wide System Assessment (TSA)
Tertiary English Language Test (TELT)
Test in Practical English Proficiency (EIKEN)
Test of Chinese as a Foreign Language (TOCFL)
Test of Practical Chinese (C.Test)
Test of English as a Foreign Language Internet-Based Test (TOEFL IBT)
Test of English for International Communication (TOEIC)
Test of English Proficiency (TEPS)

Test of German as a Foreign Language (TestDaF)
Test of the Skills in English Language (TOSEL)
TestDaF-Niveaustufen (TDNs)
Testlet response theory (TRT)
TESTMAN = Test management
TOEFL Computer Based Test (TOEFL CBT)
TOEFL Paper-Based Test (TOEFL PBT)
Tucker Lewis Index (TLI; also known as Non-Normed Fit Index)
TxtCOM = text comprehension reading ability
Unitary competence hypothesis (UCH)
United Arab Emirates (UAE)
University Grants Committee (UGC)
University of Tehran English Proficiency Test (UTEPT)
VOC = Vocabulary
Vocabulary Size Test (VST)
Weighted least squares means and variance adjusted (WLSMV)
Weighted root mean square residual (WRMR)
Youth Chinese Test (YCT)

INTRODUCTION

VAHID ARYADOUST
NATIONAL UNIVERSITY OF SINGAPORE, SINGAPORE
JANNA FOX
CARLETON UNIVERSITY, CANADA

Over the past few decades, the field of language assessment has grown in importance, sophistication, and scope. The increasing internationalization of educational and work contexts, heightened global understanding of the role of assessment in learning (e.g., Black & Wiliam, 2001; Fox, 2014; Rea-Dickens, 2001), greater emphasis on the assessment of educational outcomes (e.g., Biggs & Tang, 2007), and the concomitant expansion of the language testing industry (e.g., Alderson, 2009) have led to unprecedented changes in assessment practices and approaches. These advancements, spurred on by technological innovation and a burgeoning array of new data analysis techniques, have prompted some to suggest (e.g., McNamara, 2014) that language assessment is on the verge of a revolution.

Situated within the field of applied linguistics, modern language assessment is intensely interdisciplinary—informed not only by the broad array of sub-disciplines that comprise applied linguistics, but also by disciplines such as psychometrics, education, and psychology to name but a few. Such rich interdisciplinarity has directly contributed to theory and knowledge generation, and the development of new assessment and measurement approaches within language assessment. On the one hand, the increasing influence of sociocultural theory (McNamara & Roever, 2006) has led to renewed considerations of context (McNamara, 2007), *ecological* approaches to assessment (Fox, 2003), and a heightened awareness of the role of testing policy and practice in identity construction and power relationships (Shohamy, 2007).

In addition, drawing on cognitive theories originating within psychology, informed by advances in psycholinguistics, and supported by psychometrics

and technological innovation, language assessment researchers have been able to advance the field from multiple directions. For example, language assessment researchers are now assessing real-time, interactive language at levels of complexity and detail that only a decade ago would have been impossible. Researchers have also expanded the role of computer adaptive testing (Van der Linden & Glas, 2000); explored the performance of automatic essay evaluation (Shermis, Burstein, & Bursky, 2013) as well as natural language processing (Chapelle & Chung, 2010); extended the understanding of language constructs through advances in computational linguistics and corpus analysis (Aryadoust, 2015); and mined test data for increasingly reliable modeling, prediction, and classification (Aryadoust & Goh, 2014).

Although the literature in language assessment suggests that scholars in all parts of the world are actively engaged in research, two regions have arguably been underrepresented in this scholarly work, namely, the Pacific Rim (to some extent), and the Middle East (to a large extent). The present volume, *Trends in language assessment research and practice: The view from the Middle East and the Pacific Rim*, provides a unique overview of contemporary research in language assessment in these two regions. We have divided the 26 chapters that comprise this volume into the six following sections:

(1) Trends in data mining, psychometrics, and technology
(2) Trends in understanding individual variation in assessment outcomes: Test takers and raters
(3) Trends in diagnostic assessment: Its role in supporting learning and academic success
(4) Trends in evolving constructs of language proficiency
(5) Trends in considerations of impact and washback
(6) Trends in language assessment: Commentaries on future directions

Below, we provide a brief overview of each of these sections.

1. Trends in Data Mining, Psychometrics, and Technology

The chapters in this section examine new trends in language assessment that are informed by advances in nonlinear data mining (e.g., artificial neural networks (ANNs) and genetic algorithm-based symbolic regression), psychometrics (e.g., multidimensional Rasch measurement), and technology (e.g., functional magnetic resonance imaging (fMRI) and Automated Essay Scoring (AES)).

The first two chapters in this section investigate the potential of nonlinear modeling in language assessment. Hofstadter (1986) wrote that physics theories relied heavily on linear modeling. He stated that "Nonlinear mathematical phenomena are much less well understood than linear ones, which is why a good mathematical description of turbulence has eluded physicists for a long time, and would be a fundamental breakthrough" (Hofstadter, 1986, p. 365). Nowadays, the language assessment field is also suffering from under-recognition of the potential of nonlinear mathematical modeling. In the first chapter, Vahid Aryadoust, Mehrasa Alizadeh, and Parisa Mehran report on the results of the application of an ANN—a non-linear data mining approach—to a set of reading comprehension tests administered to Iranian applicants to graduate schools. The authors compute the linguistic features of the test items using Coh-Metrix, a computational linguistics instrument that measures various features of English texts. The authors then apply an ANN and identify the linguistic features that determine item difficulty. Aryadoust et al. discuss the implications of the study for language assessment and highlight the efficacy of nonlinear models in classification and prediction.

In the next chapter, Vahid Aryadoust extends the application of nonlinear modeling by using genetic algorithm-based symbolic regression to predict writing ability from linguistic features of texts written by tertiary-level students in Singapore. Genetic algorithm-based symbolic regression is an innovative data mining approach which has been conceptualized on the basis of evolutionary mechanisms in nature. As in the previous chapter, Aryadoust estimates the linguistic features of texts using Coh-Metrix and uses genetic algorithm-based symbolic regression to predict writing quality, as measured by raters. He discusses the application of this technique in language assessment.

In the following chapters, the authors investigate the contributions of three emerging research areas made possible by advances in psychometrics and technology. First, Purya Baghaei discusses the application of three Rasch models (i.e., unidimensional, multidimensional, and bifactor) to investigate the multidimensionality of a high-stakes reading comprehension test administered in Iran. Baghaei's findings suggest that the bifactor model provides the best fit to the data. He discusses the implications of this finding with regard to the use of bifactor models in language testing, and their potential in investigating the structure of language constructs. Baghaei explains how increasingly

sensitive and sophisticated psychometric approaches allow for the detection of multidimensionality in tests.

The next chapter examines the application of new technologies in language assessment. Kinnie Chan and Trevor Bond report on their investigation of the potential impact of computer versus human scoring in the assessment of Hong Kong students' English essays. Using many-facet Rasch measurement (MFRM), they argue that the consistency of the computer scoring of the students' essays by the Lexile Analyzer, an Automated Essay Scoring (AES) system, suggests it could be helpful in diminishing the marking stress experienced by Hong Kong language teachers, by reducing the number of essays the teachers are required to mark.

Finally, Janna Fox and Masako Hirotani explore the potential of fMRI in assessing the incremental development of language proficiency. They report on changes in the cortical activation of some Japanese learners of English, who undertook 75 hours of language instruction using the Rosetta Stone program for speaking. Measured at intervals over the 8 weeks of instruction, none of these learners' scores increased on tests of speaking proficiency. However, raters reported subtle and meaningful changes in some of the students' speech, which nonetheless failed to contribute to increased proficiency scores. Fox and Hirotani discuss implications of their study for theories of second language acquisition and language assessment.

2. Trends in Understanding Individual Variation in Assessment Outcomes: Test Takers and Raters

One of the major research topics in language assessment is construct relevant versus construct irrelevant variation in test takers' and raters' test performance and evaluation (Aryadoust & Liu, 2015). This section includes four chapters which examine factors that predict test performance and attributes which differentiate rater types. The first chapter in this section, by Aek Phakiti and Nick Zhiwei Bi, focuses on Chinese English as a Foreign Language (EFL) university students' test taking strategies. Drawing on Purpura's (2004, 2013) grammatical assessment framework and applying confirmatory factor analysis (CFA), Phakiti and Bi develop and provide initial validation evidence for two Likert-scale, lexico-grammatical strategy use questionnaires, (one trait and one state), each with Chinese translation. The questionnaires follow Purpura's advice to

explore the lexico-grammatical strategies that test takers employ while completing a form-meaning grammar test in order to better understand test performance.

In the next chapter, Limei Zhang also examines test taker performance on an English as a foreign language (EFL) reading comprehension test, investigating the relationship between college-level Chinese test takers' performance and their reported cognitive and metacognitive strategy use. Her results highlight the synergy of test takers' cognitive and metacognitive strategy use in enhancing reading test performance. Zhang's findings are considered in relation to potential pedagogical and test development practices.

Next, Purya Baghaei, Doreen Bensch, and Matthias Ziegler investigate the role of gender and major in performance on the University of Tehran English Proficiency (UTEP) Test. Applying factor analysis to examine factor structure and factor differentiation in second language proficiency on the UTEP, their findings suggest that test taker performance is not associated with examinees' major and gender. They also find psychometric evidence for the separate factors of grammar, vocabulary, and reading, which supports the reporting of separate scores for each skill.

In the last chapter in this section, Edward Schaefer highlights important issues arising in the assessment of writing when there are cultural or linguistic factors that affect raters' decisions. Investigating rater bias in the scoring of English essays written by university students in Japan, Schaefer applies MFRM and cluster analysis in tracing bias to different rater types among a group of native English-speaking raters. The application of cluster analysis in classifying raters provides a novel approach to identifying the patterns associated with rater performance.

3. Trends in Diagnostic Assessment: Its Role in Supporting Learning and Academic Success

Language assessment research has tended to focus on high-stakes examinations and undervalued in-class, continuous, and diagnostic testing (Davidson, 2010). We have identified classroom and diagnostic assessment as an important theme in research across the Middle East and the Pacific Rim.

Several of the chapters in this section discuss the trends, challenges, and potential of post-admission diagnostic assessment. Ute Knoch and Catherine Elder examine the diagnostic properties of post-admission English language assessments in Australian universities and ask the question, "Just how diagnostic are they?" Their chapter provides a framework and introduction for the other chapters in this section which discuss specific diagnostic assessment initiatives. Applying their validity framework (Knoch & Elder, 2013) in the evaluation of two post-admission diagnostic assessments, they provide evidence that neither assessment fully satisfies the validity criteria, and discuss the complex challenges that diagnostic assessment must address if it is to live up to its promised potential.

In the following chapter, Alan Urmston and Michelle Raquel reinforce the concerns raised by Knoch and Elder in their discussion of the development and implementation of a post-entry diagnostic English language assessment for undergraduate students in Hong Kong universities. Urmston and Raquel describe the complex process of test development and the many challenges (and opportunities) that have occurred as part of the development process. They discuss implications for other diagnostic assessment initiatives.

Next, Janna Fox, Janet von Randow, and Alex Volkov describe the development of a diagnostic rating scale that expands a generic diagnostic scale for assessing writing which is used as part of the Diagnostic English Language Needs Assessment (DELNA), developed at the University of Auckland in New Zealand, in order to take into account the specialized context of undergraduate engineering. Fox et al. emphasize the potential of diagnostic assessment to address the needs of undergraduate engineering students by taking into account the specific demands of this disciplinary context.

As the chapters in this section indicate, much of the potential of diagnostic assessment depends on the quality of the diagnostic measure itself. In the final chapter, Irina Eigort and Averil Coxhead investigate the Vocabulary Size Test (VST) as a diagnostic tool, describing its development, application, and use in a number of different countries within the Pacific Rim. They discuss both monolingual and bilingual versions of the test and identify the challenges arising from using different test versions in differing contexts. Their research has far-reaching

implications as they provide evidence of how test formats and contextual factors can influence the results of a diagnostic assessment procedure.

4. Trends in Evolving Constructs of Language Proficiency

The chapters included in this section explore changing definitions of language constructs which are influenced by local, contextual, and cultural concerns. The opening chapter is contributed by Roger Nunn and John Langille, who discuss models of academic communicative competence and their application in language classrooms in the United Arab Emirates (UAE). They argue that many of the academic literacy constructs, which are operationalized by external proficiency tests such as the Test of English as a Foreign Language (TOEFL), do not adequately represent the skills that UAE students need in order to meet the demands of academic study in the Middle East. Their chapter provides a useful framework for and introduction to the other chapters in this section.

In a similar vein, Dongbo Zhang, Li Li, and Shouhui Zhao identify shortcomings in research on the assessment of writing in Chinese as a second language (CSL). They focus their review on three areas: (1) the characteristics and reliability of raters; (2) issues related to test formats; and (3) the automatic rating of essays. They propose new research areas to improve the validity of assessment of L2 writing in Chinese as a second language.

In the following chapter, Siew Mei Wu and Susan Tan report on the design, development, and validation of a rubric to assess a multimodal presentation task in the context of an environmental science module in Singapore. They argue that the construct of academic literacy has been reconceptionalized to encompass multimodality; and communication and meaning making are no longer exclusively textual or verbal, but also visual, involving multimedia. Their chapter responds to the need for an assessment rubric which will allow teachers to document students' increasing awareness and control of multimodal content in performance based tasks.

Next, Mehdi Riazi and Jill Murray report on research regarding the characteristics of writing assignments in Australian universities. They discuss their findings in relation to the assessment of writing on proficiency tests and identify important differences. They suggest that these differences may well pose a threat to the validity of inferences drawn

from the proficiency tests (as a result of construct underrepresentation) and argue for more research in this under-examined area.

In the final chapter in this section, Aaron Batty and Jeffrey Stewart examine the issue of rater subjectivity occurring in the context of traditional, human-rated tests of speaking in Japan. To remedy this subjectivity, they propose a novel speaking proficiency construct, operationalized by a computer-administered speaking test, the Objective Communicative Speaking Test (OCST). Their exploratory study suggests the enhanced reliability of the OCST as a measure of speaking proficiency, and its potential as a computer-adaptive test.

5. Trends in Considerations of Impact and Washback

This section examines the consequences of tests and testing practices across the Pacific Rim and the Middle East at both societal (impact) and local classroom (washback) levels. At the beginning of section five, Atta Gebril, Deena Boraie and Elizabeth Arrigoni, working with a mandate to provide training for English language educators in Egypt via a cascade model, discuss the impact of their training approach, which was designed to increase the teachers' assessment literacy. The training approach became a catalyst not only for increasing the teachers' understanding of assessment, but also for building a professionalized community of practice. Their chapter provides an important model of the potential impact of this professional development approach.

This chapter is followed by Elvis Wagner and Santoi Wagner's investigation of the spoken texts used on three high-stakes tests of second language (L2) listening: the College English Test (CET) of China, the Test in Practical English Proficiency (EIKEN) of Japan, and the General English Proficiency Test (GEPT) of Taiwan. Their findings suggest that the lack of authenticity of the tasks and performances on these tests may undermine the potential positive washback of the tests in supporting communicative language teaching approaches.

Similarly, in the following chapter, Cynthia Wiseman considers the washback of the battery of standardized summative assessments administered in elementary, middle, and secondary schools in three Chinese provinces. By collecting teachers' accounts of the role these tests play in their teaching, she documents the ways in which the tests may