Castledown

OPEN ACCESS

Language Education & Assessment
ISSN 2209-3591
https://journals.castledown-publishers.com/lea/

Language Education & Assessment, 1 (1), 3-8 (2018)
https://dx.doi.org/10.29140/lea.v1n1.35

Then and Now: Themes in Language Assessment Research

LIA PLAKANS [a]

[a] *The University of Iowa, USA*
Email: lia-plakans@uiowa.edu

Abstract

This short reflection considers major themes in language assessment research in two years across a decade, 2007 and 2017. The themes were selected after reviewing three major journals in language testing. These themes include articles related to validity, performance assessment, classroom assessment, and technology. Each theme will be described and discussed along with comparisons between the two years. The article concludes with potential future directions in language assessment research.

Keywords: research themes, language assessment

Introduction

The practice of testing language has a very long history around the world (Spolsky, 1995). However, in the past thirty years or so, scholarship around language assessment has become a formidable presence in fields related to language acquisition, linguistics, and learning. With an inaugural issue of a new journal in language assessment and education, it seems appropriate to reflect on this work. In this short piece, I will compare topics in language assessment research between two years a decade apart, to consider common themes in the field over time.

The specific years were chosen to represent the most recent decade of language assessment scholarship, 2007 and 2017. This ten-year scope offers a conveniently round number, but it also has personal relevance as it marks the time since I completed my doctoral studies. These ten years hold professional significance to me in my shift from a student and teacher of language learning to a scholar in language testing. Due to this personal relevance, this review will include my reflections intertwined with themes in language assessment scholarship in the past decade.

Data Availability Statement: *All relevant data are within this paper.*

Approaching the Comparison

Given my involvement in language testing from 2007 to 2017, individual reflection should be imminent. However, as I wrote, I was dissatisfied that my impressions seemed not only subjective, but sadly vulnerable to memory lapses. More comfortable with published record, I turned to research journals in language assessment as representative of the body of research in the field. My attention centered on three major journals in language assessment: *Language Testing, Language Assessment Quarterly,* and *Assessing Writing*, from which I gathered full-length research articles published in the two years. The table below gives the total number of articles for each year; all three journals have increased their number of annual articles, perhaps due to digital cost-savings by publishers. I reviewed a total of 85 articles.

Table 1 *Journal Articles Reviewed in the Current Study*

Journal Name	2007	2017
Language Testing	12	18
Language Assessment Quarterly	11	19
Assessing Writing	11	14

The articles were coded for a general and specific topic within the field of language assessment. For example, a number of studies were on topics related to scoring assessments, such as rater training or rubric development. The articles were then grouped in terms of these general topics and considered in a cross-year comparison (2007 and 2017). From this coding, I selected the four topics for which more than ten articles were coded and that included articles from both years. Articles that are not included in the review either appeared only in one year or were less prevalent than the "top four." There are limitations to this approach; the process was not exhaustive—neither a research synthesis nor a meta-analysis. The coding was completed by one researcher, me. I acknowledge my interpretive bias in this reflection, but, as a witness of this time period, I am also a "data source." My professional narrative influences the way in which the trends in the field were both identified and interpreted, which is a limitation, but also the point of a reflection. In the following pages, I present the common themes found as well as the shifts between the two years.

Themes Across the Decade

The review of journals revealed the robust nature of the field of language assessment. The increasing number of publications from 2007 to 2017 gives a sense that this is a flourishing discipline (or sub-discipline). Many topics within language testing appeared during these years, some only once, but others were clearly of major interest to a number of scholars. These perennial topics will be the focus of this section: validity (16 articles), performance assessment (13 articles), classroom assessment (11 articles), and technology (15 articles).

Certainly, the regular appearance of articles related to validity is not a surprise. As a foundational aspect of test quality and in its centrality to test design, validity is a core concept in language assessment. In another ten years, validity will most likely still be a topic present in journals of language assessment research. The validity studies published in the two years reflect theoretical shifts in thinking about validity in assessment. In 2007, the field had moved away from the "tool kit" approach of gathering many parallel kinds of assessment (Chapelle, 1999). Instead, during this time, a unified view of validity appeared in research. Studies of validity looked at construct validity across tests, such as in writing assessment or in rating scales (Beck & Jeffrey, 2007; Sawaki, 2007). With this backdrop of construct validity, several studies focused on "validating" measures (Llosa, 2007; Reinheimer, 2007). Spanning ahead a decade to 2017, construct validity still appears to be a backbone for validity research in language assessment. In these recent articles, frameworks and models are used to consider validity in assessments for primary school leaners (Tengberg, 2017) and for workplace English (Yoo & Manna, 2017). While the idea of construct validity is foundational in assessment, it has been critiqued for not being concrete enough for clear application; yet, it will likely continue to maintain its place as a dominant approach to thinking about language and how we evaluate tests. The articles in 2017, however, reflect an approach to validity beyond construct, moving to an argumentation approach to validity (Chapelle, Engright, & Jamieson, 2008; Kane, 2012). An argument-based system for validity investigation has been proposed as a useful and comprehensive approach based on logic models (Bachman & Palmer, 2010). A number of studies consider evidence for inferences made in validity arguments and the interpretations of scores (Kyle & Crossley, 2017; LaFair & Staples, 2017; Plakans & Gebril, 2017). It will be interesting to see if this view still appears in scholarship in ten years, 2027, and how it might evolve further.

Another topic that appeared frequently in language assessment research in 2007 and 2017 is performance assessment. The practice of assessing language use directly through performances emerged in the 1970s and appears to be a mainstay in research. In both years, scholars wrote on the use of portfolios to evaluate language (Dysthe, Engelsen, & Lima, 2007; Lam, 2017); however, the type of performance task appeared less than the critical issues of rating performances through scales, rubrics, and raters. In 2007, research on scales focused on different formats, such as analytic and holistic scales (Barkaoui, 2007; Knoch, 2007; Xi, 2007). In 2017, researchers asked questions about the content and contexts of scales, such as a scale designed to assess functional adequacy in writing (Kuiken & Vedder, 2017) or a rubric with criteria for assessing science writing (Rakedzon & Baram-Tsabari, 2017). This might suggest the field's growing sophistication in rating, with past questions about general rubrics and current attention on specific purpose assessment. In both years, raters were a topic for investigation. Comparing studies on raters in the two years is particularly interesting as the shift seems to be from practical questions of rater training (e.g., Knoch, Read, & von Randow, 2007) to more nuanced studies of raters and their interactions with the process of rating. For example, in 2017 scholars investigated raters' perception of tasks (Wang, Engelhard, Raczynski, Song & Wolke, 2017) as well as raters' potential impact on feedback in assessment (Trace, Janssen, & Meier, 2017). Our focus in research may be shifting to the "human" elements in performance assessment rather than the reliability of the scores. Raters have a considerable role in these assessments and recognizing how they think about and enact scores will be a fruitful area for ongoing research. As a general area of research, like validity, performance assessment seems to have garnered a near permanent position in language assessment research.

In both years, a number of articles were published that centered on classroom assessment issues. Having classroom assessment as an ongoing theme in the field is significant as it speaks to teachers, students, and other stakeholders in assessment. The studies on classroom assessment tended to be context driven—Hong Kong writing classrooms, a critical thinking course, or a Chinese EFL teachers' practices (Carroll, 2007; Lee, 2007; Wang, 2017). In most cases, the focus is on the use of assessment in context, rather than studies that isolate the test from its environment. Other classroom assessment studies explored more philosophical questions regarding the connection between formative assessments for learning and the larger context of assessment culture (Barlow, Liparulo, & Reynolds, 2007; Xiao, 2017). These conversations bridge the potential divide in language testing between large scale standardized testing and assessment for classroom purposes (Inbar-Lourie, 2008; Malone, 2013). My impression, before reviewing the journals, was that these attempts of converging language testing theory in the two contexts was very recent, and I was pleased to discover that it was part of the conversation back in 2007.

Technology was a topic that appeared in language testing journals in 2007; ten years later, however, it is nearly a deluge. The earlier studies looked at automation of parts of test delivery, such as video use in computer-based tests (Ockey, 2007) or a multimedia assisted test (Lee, 2007). As technology use has increased and become more sophisticated, so has its application in language testing. Two major areas of technology use in testing emerged from the 2017 journal articles: automated scoring and corpus linguistic research. Automated scoring has been a major issue particularly in writing assessment as the medium of written text lends itself readily to computer language processing. Several studies in 2017 investigated how automated writing assessment could be used in formative contexts (Bridgeman & Ramineni, 2017; Wilson, Roscoe, & Ahmend, 2017), which shows a desire to apply this technology beyond its more prevalent use in large scale standardized assessments. A special issue in 2017 of *Language Testing* (edited by Cushing, 2017) explores the potential intersection with corpus linguistics, which uses technology to analyze large text data sets, and language testing. In addition to scoring and corpus use, several recent studies investigated test taker issues with technology in language testing, such as keyboarding and computer literacy (Jin & Yan, 2017; Ling, 2017). These recent investigations of user fidelity and comfort with computers is surprising as such questions should have been asked and, ideally, answered in the early days of computer-based testing. While application of technology to language testing has been popular for much longer than ten years, the research that explores it seems to be heading in the direction of innovation, while still attending to highly practical matters.

In addition to these four major themes, some other patterns emerged in the review. Sophisticated approaches to analysis drawn from educational measurement and applied to language testing were found in both years. For 2007, differential item functioning (DIF) seemed to have captured the attention of the field, resulting in a special issue of *Language Testing.* In 2017, an analysis featured in several studies was generalizability theory (G-theory). Another observable shift was more attention in research on issues around tests, such as score use, test preparation, and test takers' perceptions (Gebril & Eid, 2017; Khabbazbashi, 2017). This suggests a broader lens beyond tests and instruments to the complex contextual aspects that surround testing.

Looking Ahead

Completing this review of articles from these two years across a decade, hints at the future and what research topics might appear in journals ten years from now. The growing interest in issues around tests is likely to continue, with stakeholders, test takers, and test users, garnering interest. Along these lines, the area of policy seems ripe for investigation as language policy as well as other policies (e.g., citizenship) often utilize language assessments. Certainly, the field of language assessment has contemplated policies, particularly with ethical practice and fairness surveys. In order to investigate these domains, different approaches to research, such as narrative inquiry, ethnography, and critical lenses will be necessary. Along with more attention to social and user-focused research and methods, the major themes explored in this reflection will continue to appear—validity, performance assessment, classroom assessment, and technology. Reflecting on this snapshot of one decade, it is unequivocal that the field of language assessment has and will continue to offer useful, valuable, and insightful research for many years to come.

References

Bachman, L., & Palmer, A. (2010). *Language assessment in practice.* Oxford, UK: Oxford University Press.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12* (2), 86-107. https://dx.doi.org/10.1016/j.asw.2007.07.001

Barlow, L., Liparulo, S. P., & Reynolds, D. W. (2007). Keeping assessment local: The case for accountability through formative assessment. *Assessing Writing, 12* (1), 44-59.

https://dx.doi.org/10.1016/j.asw.2007.04.002

Beck, S. W., & Jeffery, J. V. (2007). Genres of high-stakes writing assessments and the construct of writing competence. *Assessing Writing, 12* (1), 60-79. https://dx.doi.org/10.1016/j.asw.2007.05.001

Carroll, D. W. (2007). Patterns of student writing in a critical thinking course: A quantitative analysis. *Assessing Writing, 12* (3), 213-227. https://dx.doi.org/10.1016/j.asw.2008.02.001

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

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

Cushing, S. T. (2017). Corpus linguistics in language testing research. *Language Testing, 34,* 441-449. https://dx.doi.org/10.1177/0265532217713044.

Gebril, A., & Eid, M. (2017). Test preparation beliefs and practices in a high-stakes context: A teacher's perspective. *Language Assessment Quarterly, 14*(4), 360-379. https://dx.doi.org/10.1080/15434303.2017.1353607

Inbar-Lourie, O. (2013). Guest editorial to the special issue on language assessment literacy. *Language Testing, 30,* 301-307.

Kane, M. (2012). Validating score interpretations and use. *Language Testing, 29,* 3-17.

Khabbazbashi, N. (2017). Topic and background knowledge effects on performance in speaking assessment. *Language Testing, 34* (1), 23-48. https://dx.doi.org/10.1177/0265532215595666

Knoch, U. (2007). 'Little coherence, considerable strain for reader': A comparison between two rating scales for the assessment of coherence. *Assessing Writing, 12* (2), 108-128. https://dx.doi.org/10.1016/j.asw.2007.07.002

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12* (1), 26-43. https://dx.doi.org/10.1016/j.asw.2007.04.001

Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing, 34* (4), 513-535. https://dx.doi.org/10.1177/0265532217712554

Kuiken, F., & Vedder, I. (2017). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing, 34* (3), 321-336. https://dx.doi.org/10.1177/0265532216663991

Lam, R. (2017). Taking stock of portfolio assessment scholarship: From research to practice. *Assessing Writing, 31*, 84-97. https://dx.doi.org/10.1016/j.asw.2016.08.003

Lee, I. (2007). Feedback in Hong Kong secondary writing classrooms: Assessment for learning or assessment of learning? *Assessing Writing, 12* (3), 180-198. doi: 10.1016/j.asw.2008.02.003

Llosa, L. (2007). Validating a standards-based classroom assessment of English proficiency: A multitrait-multimethod approach. *Language Testing, 24* (4), 489-515.

Malone, M. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing, 30,* 329-334.

Plakans, L., & Gebril, A. (2017). Exploring the relationship of organization and connection with scores in integrated writing assessment. *Assessing Writing, 31*, 98-112. https://dx.doi.org/10.1016/j.asw.2016.08.005

LaFlair, G. T., & Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference in the validity argument for a high-stakes speaking assessment. *Language Testing, 34* (4), 451-475. https://dx.doi.org/10.1177/0265532217713951

Rakedzon, T., & Baram-Tsabari, A. (2017). To make a long story short: A rubric for assessing graduate students' academic and popular science writing skills. *Assessing Writing, 32*, 28-42. https://dx.doi.org/10.1016/j.asw.2016.12.004

Reinheimer, D. A. (2007). Validating placement: Local means, multiple measures. *Assessing Writing, 12* (3), 170-179. https://dx.doi.org/10.1016/j.asw.2008.02.004

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

Tengberg, M. (2017). National reading tests in Denmark, Norway, and Sweden: A comparison of construct definitions, cognitive targets, and response formats. *Language Testing, 34* (1), 83-100. https://dx.doi.org/10.1177/0265532215609392

Trace, J., Janssen, G., & Meier, V. (2017). Measuring the impact of rater negotiation in writing performance assessment. *Language Testing, 34* (1), 3-22. https://dx.doi.org/10.1177/0265532215594830

Yoo, H., & Manna, V. F. (2017). Measuring English language workplace proficiency across subgroups: Using CFA models to validate test score interpretation. *Language Testing, 34* (1), 101-126. https://dx.doi.org/10.1177/0265532215618987

Walters, F. S. (2007). A conversation-analytic hermeneutic rating protocol to assess L2 oral pragmatic competence. *Language Testing, 24* (2), 155-183.

Wang, J., Jr. Engelhard, G., Raczynski, K., Song, T., & Wolfe, E. W. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing, 33*, 36-47. https://dx.doi.org/10.1016/j.asw.2017.03.003

Xi, X. (2007). Evaluating analytic scoring for the TOEFL® Academic Speaking Test (TAST) for operational use. *Language Testing, 24* (2), 251-286.

Xiao, Y. (2017). Formative assessment in a test-dominated context: How test practice can become more productive. *Language Assessment Quarterly, 14* (4), 295-311. https://dx.doi.org/10.1080/15434303.2017.1347789

Author Biodata

Lia Plakans is associate professor of Foreign Language and ESL Education at the University of Iowa. Her research focuses on second language learning with particular emphasis on language assessment and literacy. She has directed assessment research grants funded by Educational Testing Service (ETS), Cambridge Michigan Language Assessments, and Language Learning journal. She is an associate editor for Language Assessment Quarterly. She has co-authored the books Assessment Myths: Applying Second Language Research to Classroom Teaching with University of Michigan Press. She was an English language teacher for over 15 years in Iowa, Texas, Ohio, and Latvia.