Emergent Trends and Research Topics in Language Testing and Assessment

Tuğba Elif Toprak Yıldız [1]

Abstract: This study, which is of descriptive nature, aims to explore the emergent trends and research topics in language testing and assessment that have attracted increasing attention of language testing and assessment researchers. To this end, 300 articles published within the last seven years (2012-2018) in two leading journals of the field were analyzed by using thematic analysis method. Overall, the results demonstrated that the assessment of language skills still constitute the backbone of language testing and assessment research. While the term communicative has become the established norm in language testing and assessment, the field has grown more interested in professionalization, understanding the dynamics that underlie test performance and validation. Moreover, the results revealed that even though the latest advancements in the fields of computer, cognitive sciences and information/communication technologies seem to make their way into language testing and assessment, more research is needed to make the most of these advancements and keep up with the rapidly changing nature of communication and literacy in the 21st century. The results are discussed and the implications are made.

Keywords: Language testing; language assessment; emergent trends in language testing; educational assessment; educational testing

DOI: 10.29329/mjer.2019.202.4

[1] **Tuğba Elif Toprak Yıldız,** Dr. Öğr. Üyesi, İzmir Bakırrçay Üniversitesi, Yabancı Diller Yüksekokulu, İzmir, Türkiye. ORCID: 0000-0003-0341-229X

İrtibat Yazarı: toprak@bakircay.edu.tr

INTRODUCTION

The field of language testing and assessment has approximately 2000 years of history dating back to the Han Dynasty which conducted assessment and testing practices to choose men for several missions (Spolsky, 2008). However, the field emerged as a sub-discipline of applied linguistics only a few decades ago and gradually turned into not only a vibrant area of research but also an industry, mainly as a consequence of a pursuit of reliability and a market-driven demand for foreign/second language proficiency testing. Furthermore, language testing and assessment has been highly affected by the developments taking place in areas such as psychometrics, educational assessment, and linguistics, and reflected the trends and paradigms prevalent in these areas to a considerable extent. Historically, during the period between the emergence of Chinese civil service examinations and the 19th century, language testing and assessment was characterized by open-ended and oral examinations that lacked a solid linguistic and measurement theory. Nevertheless, in the 20th century, which was called the psychometric-structuralist era among language assessment community, the field became increasingly concerned with the objective and reliable methods of testing and the tenets of structural linguistics and classical test theory were widely incorporated into language testing and assessment practices (Stansfield, 2008).

During and the aftermath of two world wars, the agenda of the 40s, 50s as well as the 60s was heavily loaded with issues such as the launch of military programs for improving and measuring oral language abilities of the troops, aptitude testing and discrete-point testing in which the primary concern is psychometric reliability (Lado, 1961). However, the 1980s witnessed a paradigm shift, a shift from a more psychometric-oriented, structural and behaviorist view to a more communicative and integrated one, particularly motivated by the work of renowned applied linguistics such as Savignon (1972), Canale and Swain (1981) and

Widdowson (1983). While the 90s hardened the cement of the communicative view of language testing and assessment, novel approaches, trends and research topics continued to make their way into language testing and assessment agenda during the 2000s.

Within the last two decades, language testing and assessment research has witnessed the adoption, adaptation, and refinement of a wide array of methods and means for research while various topics and issues have been and are still being explored in a great number of studies. In his overview on language testing and assessment, Bachman (2000) predicted that future language testing and assessment research would focus on professionalization of the field and validation. More recently, Harding (2014) eloquently provided an overview on the current issues and future research avenues in Communicative Language Testing and argued that rather than fading away gracefully, CLT has become the implicit orientation or dominant paradigm in modern language testing. Apart from offering crucial and useful insights into the term communicative and the main issues related to CLT testing,

Harding's article was significant in that it focused on the issue of adaptability itself and related future research avenues. These research directions, according to Harding (2014), would be adaptability to i) understand and deal with different varieties of English, ii) understand and use appropriate pragmatics, iii) employ corpus approaches to ensure predictability and modeling in language test tasks, and iv) use novel test tasks that employ 21st century communication platforms and tools.

The present study aims to explore major research topics that are prevalent in the field and highlight significant trends as represented by articles published in two leading journals of the field, Language Testing and Language Assessment Quarterly. It is hoped that the implications of the present study would be useful to language testing and assessment researchers, graduate students concentrating on language testing and assessment, and their academic advisors.

METHOD

This section presents detailed information about the research design of the study, journal selection process, the properties of the journals and the corpus, and finally data analysis procedures.

Research design

The present study employed the descriptive research design, which can be used to i) describe the features of a given population or area at hand, ii) provide a detailed and accurate account of the features of individuals, situation and groups, and iii) display the characteristics of individuals, situations and groups and the frequency of the occurrence of the phenomena (Dulock, 1993).

Journal selection and the corpus

In order to determine scholarly work on language testing and assessment to be scrutinized, the researcher communicated with five scholars working on language testing and assessment and asked them to name research outlet/s in the field which i) is/are directed to an international audience; ii) they followed to be knowledgeable about the studies conducted language testing and assessment ii) they thought was/were reflecting the trends in language testing and assessment. All experts unanimously agreed that Language Testing and Language Assessment Quarterly were leading journals that are dedicated solely to language testing and assessment and maintained that these two journals not only reflected the current state-of-the-art in the field but also led it. Hence, the articles comprising the corpus were selected from these two journals. These articles were published between the years 2012-2018, therefore covering a seven-year-period. The corpus included 300 articles; 166 articles published in Language Testing and 134 articles published in Language Assessment Quarterly. Both journals are published quarterly, and overall, each volume features a special issue focusing on a specific topic. Detailed information about the corpus is presented in Table 1.

Table 1. Detailed information about the corpus

Language Assessment Quarterly	Year	Issue 1	Issue 2	Issue 3	Issue 4
	2012	4*	5	5	4
	2013	6*	5	5	5
	2014	4*	7*	5	3
	2015	7*	3	3	4
	2016	4	2	6	7
	2017	4	5*	5*	7
	2018	9*	5	5*	-
Language Testing	Year	Issue 1	Issue 2	Issue 3	Issue 4
	2012	5	7	6	6
	2013	6	6	6*	6
	2014	6	5	5*	6
	2015	6	6	6*	6
	2016	6	7*	6	6
	2017	6	6	6	7*
	2018	6	6	4*	6

Note: Special issues are indicated by asterisks.

Data analysis

The present study utilized thematic analysis (Braun & Clarke, 2006) with an inductive approach in which coding and theme development were directed by the content of the journal articles. Braun and Clarke (2006, p.79) define thematic analysis as "a method for identifying, analyzing and reporting patterns (themes) within data." Initially, the title, abstract, methods and results sections of each article were examined in detail to determine research topics that it addressed. On most occasions, the articles concentrated on several research topics rather than dealing with one issue. After all topics were listed, a more condensed and compact categorization was created under the banner of themes. To illustrate, four topics emerged from Barkaoui's (2014) study entitled "Examining the impact of L2 proficiency and keyboarding skills on scores on TOEFL-iBT writing tasks". These topics were "writing assessment", "computer-based assessment", "keyboarding skills" and "TOEFL-iBT writing". The topics were subsumed under themes as "language skills assessment", "assessment types", "stakeholder characteristics" and "international standardized tests" respectively. The frequencies of these topics/emerging themes were calculated. The researcher conducted the analyses at four different intervals to ensure the reliability of the codings.

Results

This section reports on the results of the thematic analysis. The present study sought to determine the topics that have enticed the attention of language testing and assessment researchers within the last seven years. The frequencies of the research topics and their corresponding themes are presented in Table 2 and Table 3 respectively.

Table 2. Topics addressed in language testing and assessment research

Topics	Frequency
L2 speaking assessment	45
Test development/validation	35
Rater characteristics	34
Standard setting/alignment to the CEFR/assessment policy	27
TOEFL	27
L2 writing assessment	25
L2 reading assessment	24
L2 listening assessment	19
Assessment of lexical competence	18
EAP/ESP assessment	18
Diagnostic testing	17
Assessment literacy	16
Test taker characteristics	14
Assessment of young language learners	12
Validity	11
Assessment of other languages	11
Statistical methods	10
Proficiency testing	7
Assessment in Canada	7
Assessment of grammatical competence	7
Reliability	7
Differential item functioning	6
Placement testing	6
Integrated testing	6
High-stakes testing	6
Web/computer-based testing	5
Assessment in China	5
Assessment in Japan	5
Automated scoring	4
Assessment in Taiwan	4
TOEIC	4
German language assessment	4
Chinese language assessment	4
Dutch language assessment	4
Dynamic assessment	3
Assessment of pragmatic competence	3
Japanese language assessment	3
British/American sign language assessment	3
IELTS	2
Fairness	2
Assessment of bilinguals	2
Eye-tracking methodology	2
Natural language processing	2
Teacher perceptions	2
French language assessment	2
World Englishes	2
Classroom-based assessment	1
Performance-based assessment	1
Peer assessment	1
Task-based assessment	1
Self-assessment	1
Assessment of specific language impairment	1
Washback	1
Hebrew language assessment	1
Spanish language assessment	1
TestDaF	1

As can be seen in Table 1, 58 different topics were investigated in 300 articles published in Language Testing and Language Assessment Quarterly. The most researched topics were second language speaking assessment (45 studies), test development and validation (35 studies), rater characteristics (34 studies), test equating and standard setting practices –specifically in the context of the Common European Framework of Reference for Languages (the CEFR, European Council, 2001) (27 studies), properties of the TOEFL (Test of English as a Foreign Language) (27 studies), second language writing assessment (25 studies) and second language reading assessment (24 studies). Although these were the most addressed topics, a wide array of topics including assessment literacy among teachers, assessment of young language learners and proficiency testing was also dealt with in a great number of studies.

To capture a clearer and more unified picture of the current landscape of the field and be more able to detect the emergent research trends, the topics exhibiting similar properties were subsumed under broader categories, i.e., themes. For instance, while L2 reading, speaking, writing and listening assessment were collapsed into a theme labelled "language skills assessment", issues such as validity, reliability, fairness and washback were grouped under the theme "test properties". Table 3 presents information about these themes and their frequencies.

Table 3. Emergent themes

Themes	Frequency
Language skills assessment	138
Assessment/testing types	74
Stakeholder characteristics	50
Test development/validation	35
International standardized tests	34
Standard setting/Alignment to the CEFR/assessment policy	27
Region-specific assessment	22
Test properties	19
Statistical methods	16
Assessment literacy	16
Assessment of special groups	15
Assessment of other languages	14
Novel technologies and methods	9

Table 3 demonstrates that the assessment of language skills (138 instances) was by far the most studied issue in the field of language assessment. This classic theme was followed by assessment types (e.g., placement, proficiency, diagnostic, academic and specific purposes assessments); stakeholder characteristics (characteristics of examinees, examiners and raters); test development and validation; the properties and uses of international standardized tests (e.g., the TOEFL, IELTS, TOEIC); and test equating and standard setting practices especially in the European context. Other emerging themes were test properties (e.g., validity, reliability, practicality) and how they could be enhanced, region-

specific assessment --assessments conducted in several regions of the world (e.g., Taiwan, Canada, and China), use of various statistical methods (e.g. item response theory modelling, Rasch modelling, cognitive diagnosis psychometric modelling, differential item functioning), assessment literacy of language teachers and assessment of special groups (young/very young language learners, visually/hearing/language impaired users, bilinguals). Finally, assessments conducted in languages other than English (e.g., Dutch, Chinese, Spanish and sign languages) and the use of novel methods and technologies such as eye-tracking methodology, automated scoring and natural language processing tools were also dealt with in a relatively limited number of studies.

DISCUSSIONS AND CONCLUSIONS

In his review on language testing and assessment at the turn of the century in American Association for Applied Linguistics newsletter, Bachman (2000) presented a vivid picture of language assessment in relation to the trends in applied linguistics and provided a timeline of how the field developed and changed. According to Bachman (2000), while utmost attention was paid to developing psychometrically-rigorous language tests and examining the psychometric properties of these tests during the 60s and the 70s, the field experienced a shift in the 80s that was inspired and shaped by a more communicative view of language. At the first Language Testing Research Colloquium (LTRC) that was held in 1979, language testing and assessment emerged as a subfield of linguistics with its own research agenda and methodology and embraced the notion of communicative competence as the target of its inquiries. Moreover, during the 90s, the field witnessed drastic changes including the emergence of computer-adaptive testing (Chalhoub-Deville & Deville, 1999) and the inclusion of novel concepts such as rater training (Shohamy, Gordon, & Kraemer, 1992). Since these times, the focus of the field has been on the discoursal, sociolinguistic and pragma-linguistic features of the language. Consequently, language assessment researchers have begun to address issues such as the operationalization of communicative language ability and its assessment, uses of language tests in various situations, novel assessment methods and formats, scoring procedures and their impacts on assessment performance, validity issues and authentic language tests.

Overall, the findings of the study reveal that the interest in the assessment of communicative language ability in different contexts (e.g., academic, workplace, vocational), for different purposes (e.g., placement, admission, certification, standard-setting) and in various formats (e.g., paper-based, computer-based, face-to-face) continues to exist. This situation supports Harding's (2014) argument over the current state of communicative language testing, which he put into words "CLT still remains the theoretical construct underlying many tests and informs a focus on authentic language tasks that involve performance, but it is now accepted as a conventional approach across many testing situations" (p.194). In other words, much language testing and assessment is communicative in that

researchers base their scholarly efforts on communicative language ability theories and employ authentic test tasks that involve interaction and authenticity.

With regard to research themes, the most researched theme was the assessment of language skills; namely reading (e.g., Aryadoust & Zhang, 2016; Tengberg, 2017), writing (e.g., Ling, 2017, Kuiken & Vedder, 2017), listening (e.g., Lee & Winke, 2013; Suvorov, 2015, Wagner, 2013) and speaking. In particular, the assessment of speaking performance was a major concern in a great number of studies (e.g., Babaii, Taghaddomi, & Pashmforoosh, 2016; Cai, 2015; Hirai & Koizumi, 2013; Jin, Mak, & Zhou, 2012; Kim, 2015; Nakatsuhara, Inoue, Berry, & Galaczi, 2017; Sato, 2012). Much of the research on language assessment seems to continue to focus on the four skills (Bachman, 2000), yet the assessment of lexical and grammatical competence has also attracted considerable attention of language testing and assessment researchers. On the other hand, studies dealing with pragmatic competence remained underrepresented.

Assessment types were the second most researched theme. In these studies, researchers examined the practices of the proficiency, placement, diagnostic, integrated, peer, classroom-based EAP, ESP and dynamic assessments from various aspects. Although the majority of studies falling in this theme focused on EAP/ESP (e.g., Appel & Wood, 2016; Green & Hawkey, 2012; Pill & McNamara, 2016), diagnostic (e.g., Harding, Alderson, & Brunfaut, 2015; Li, Hunter, & Lei, 2016; Li & Suen, 2013), proficiency (e.g., Barkaoui, 2014; Cai, 2013; Denies & Janssen, 2016) and placement assessment (e.g., Eckes, 2017; Kokhan, 2013), the number of studies dealing with more alternative forms of assessment such as dynamic, classroom-based, peer, task-based and self-assessment was limited (e.g., Aryadoust, 2016; Butler & Zeng, 2014; Suzuki, 2015). Considering that in the recent decades there has been an interest in alternative approaches to assess students' performances (Chapman, 2003), this situation would be surprising. Nevertheless, taking the need for proficiency and placement testing and the high-stakes nature of these assessments into consideration, this never-diminishing interest in relatively traditional types of assessments is also understandable. Moreover, in most studies reviewed, researchers tried to determine the ways in which the potential of these relatively traditional assessments could be exploited best since these assessments are typically used to make important decisions about individuals' life such as entering a college, receiving language education at an appropriate level, becoming eligible for being a health professional, applying for citizenship and certification (e.g., Lin & Zhang, 2014; Bridgeman, Cho, & DiPietro, 2016; Choi, 2017, Farnsworth, 2013).

Third, a great number of studies focused on examining the characteristics of stakeholders, who could be defined as the examiners, examinees, assessment developers and raters. This strand of research has mainly concentrated on capturing and understanding the factors that may impact examinees' performances on test tasks. To illustrate, while from test takers' view, test takers strategy

use, background, attitudes towards a certain type of assessment, perspectives and even keyboarding skills were investigated (e.g., Murray, Riazi, & Cross, 2012; Zhang, Goh, & Kunnan, 2014), raters' accent familiarity, linguistic background, negotiation abilities, perspectives, training, experience, judgements, decision style and cognition were examined from the rater point of view (e.g., Baker, 2012; Davis, 2016; Eckes, 2012; Huang, Alegre, & Eisenberg, 2016; Kang, 2012; Winke, Gass, & Myford, 2013). The majority of the studies falling under this theme were conducted to investigate rater behaviour and performance-- ultimately to implicate factors that might influence the reliability of assessments.

Fourth, test development and validation issues were investigated in a number of studies where researchers elaborated on the development process and validation of an extensive range of tools such as morphology and reading tests for young learners (e.g., Goodwin, Huggins, Carlo, Malabonga, Kenyon, Louguit, & August, 2011), sign language tests (e.g., Bochner, Samar, Hauser, Garrison, Searls, & Sanders, 2016) rater attitude tests (e.g., Hsu, 2016), story re-telling speaking test (e.g., Hirai & Kouzimi, 2013) and intercultural pragmatics test (e.g., Timpe-Laughlin & Choi, 2017). In these scholarly efforts, the researchers opted for a program of validation either/both as a conjecture for research and as a procedure for ensuring quality in test design, creation and use. The findings demonstrated that, although at the broader level these tools targeted at primary language skills such as speaking and writing, actually, they focused on measuring more grain-sized skills and attributes such as the effects of native language definitions and cognate status of test items, and authorial voice strength in argumentative writing.

Fifth, the properties and uses of international standardized tests such as the TOEFL, TOEIC, and IELTS were investigated in a considerable number of studies. This outcome is not surprising since these tests are mostly high-stakes in nature, may function as gate-keepers, and are used to make significant decisions about individuals' lives. Accordingly, in these studies, a wide range of issues such as understanding the relationship between speaking test scores to real-life academic speaking (e.g., Brooks & Swain, 2014), using standardized test scores for placement of international undergraduate students in language courses (e.g., Kokhan, 2013¸ Papageorgiou & Cho, 2013), the differences between raters of different ethnic backgrounds in speaking assessing test tasks (e.g., Wei & Llosa, 2015), the effect of keyboard type on writing performance (e.g., Ling, 2017) were addressed. Most of the studies falling in this category specifically focused on the properties of the TOEFL.

Sixth, a number of studies focused on the issues surrounding standard-setting, equating and linking particularly within the CEFR context to promote transferability and accountability of the assessments in the European countries. To illustrate, these studies targeted at validating a particular national alignment to the CEFR (e.g., Ilc & Stopar, 2015), setting cut scores on an English placement test using the prototype group method (e.g., Eckes, 2017) and applying multifaceted Rasch

measurement in standard setting procedures (e.g., Hsieh, 2013a, 2013b). A relatively limited number of studies focused on various issues such as investigating how language assessment practices are carried out in a specific part of the world (e.g., Pan & Qian, 2017; Saida, 2017), understanding the test properties (i.e., validity, reliability, fairness, practicality) (e.g., Attali, Lewis, & Steier, 2013; Shaw & Imam, 2013) and showcasing the application of a statistical method or procedure that is not adequately known to language assessment community. Unidimensional and multidimensional item response theory models, differential item functioning, logistic regression models and generalizability theory (e.g., Fidalgo, Alavi, & Amirian, 2014; Han, 2016; Koo, Becker, Kim, 2014) were among these methods and procedures. Although statistical modelling procedures such as the generalizability theory, item response theory modelling, and structural equation modelling have been quite in use in language testing and assessment research since the 90s, these modelling procedures seem to take up much of the space in the statistical toolbox of language testers when compared to other relatively more up-to-date procedures.

Apart from these issues, evaluating language teachers' assessment literacy (e.g., Vogt & Tsagari, 2014; Lam, 2015), investigating the assessment of special groups such as bilinguals (e.g., Sanchez, Rodriguez, Soto-Huerta, Villarreal, Guerra & Flores, 2013) language-impaired individuals (e.g., Katzenberger & Meilijson, 2014) and young/very young language learners (e.g., Huang & Konold, 2014; Lee & Winke, 2018) were dealt with in a relatively limited number of studies. Finally, several studies explored the assessment of languages other than English such as German, French and Chinese (e.g., Eckes, 2014; Granfeldt & Ågren, 2014; Jin & Mak, 2012), assessment of sign languages (e.g., Haug, 2012; Mann, Roy & Morgan, 2016); and the use of novel methodologies in language assessment such as automated scoring and evaluation (e.g., Chapelle, Cotos, & Lee, 2015; Hoang & Kunnan, 2016; Xi, Higgins, Zechner, & Williamson, 2012), natural language processing tools (e.g., Kyle, Crossley, & McNamara, 2016) and eye-tracking (e.g., Suvorov, 2015; Bax, 2013).

Overall, the findings suggest that the classic themes--the assessment of language skills and the uses of several assessment types still constitute the backbone of language assessment research and scholarly interest in these issues will not cease in the future. In his comprehensive overview of the field in 2000, Bachman predicted that professionalizing the field and validation research would be vital to language testing and assessment research. The professionalization of the field, according to Bachman (2000) could be ensured through training language testing professionals and developing standards of language testing and assessment practices. The findings of the present study demonstrated that these two issues have drawn researchers' attention, as can be inferred from the number of studies focusing on raters' training, teachers' assessment literacy, and standard-setting procedures especially in the CEFR context. Furthermore, a considerable and increasing degree of attention has also been paid to understanding the impact of rater and examinee characteristics on assessment results for

enhancing the reliability of assessments and understanding the functioning of standardized tests. These actions are expected to help better account for crucial implications arising from the assessments.

Validation, the second vital issue to language testing in view of Bachman (2000), seems to also have attracted the attention of language testing and assessment researchers since the number of studies that marshal information about the factors and processes that impact test performance, employ a variety of research tools and consider the consequences of test use is great (e.g., Chapelle, Cotos, & Lee, 2015; Knoch & Chapelle, 2017). At this point, it is fair to say that the field has grown as a profession that conducts assessment practices by examining the dynamics of broader educational, societal and economic contexts. Nevertheless, it seems that more research efforts are needed to incorporate the latest advancements in computer sciences, communication and cognitive sciences into language testing and assessment. Although these advancements seem to make their way into the field, the number of studies exploiting novel approaches and tools such as fMRI, eye tracking, natural language processing, corpus approaches and digital communication platforms remained relatively scarce. Moreover, the findings demonstrate that attention that potential of alternative forms of assessments such as dynamic, peer and self-assessment for language assessment has received is quite scant. Apart from these, areas that were highlighted by Harding (2014) as fruitful research avenues deserving further attention, such as the development of tests that measure test takers' ability to handle different varieties of English, the use of corpus approaches in predicting and modeling language test tasks, the use of novel test tasks that are based on various digital written communication modes and social media tools, still seem to be in need of scholarly attention and inquiry. The findings of the present study clearly demonstrated that there is a considerable need for exploring these research avenues to keep up with the rapidly changing communication and information technologies, due to which there has been a dramatic change in the nature and meaning of communication and literacy in the 21st century.

REFERENCES

Appel, R., & Wood, D. (2016). Recurrent word combinations in EAP test-taker writing: Differences between high-and low-proficiency levels. *Language Assessment Quarterly, 13*(1), 55–71.

Aryadoust, V. (2016). Gender and academic major bias in peer assessment of oral presentations. *Language Assessment Quarterly, 13*(1), 1–24.

Aryadoust, V., & Zhang, L. (2016). Fitting the mixed Rasch model to a reading comprehension test: Exploring individual difference profiles in L2 reading. *Language Testing, 33*(4), 529–553.

Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing, 30*(1), 125–141.

Babaii, E., Taghaddomi, S., & Pashmforoosh, R. (2016). Speaking self-assessment: Mismatches between learners' and teachers' criteria. *Language Testing, 33*(3), 411–437.

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.

Baker, B. A. (2012). Individual differences in rater decision-making style: An exploratory mixed-methods study. *Language Assessment Quarterly, 9*(3), 225–248.

Barkaoui, K. (2014). Examining the impact of L2 proficiency and keyboarding skills on scores on TOEFL-iBT writing tasks. *Language Testing, 31*(2), 241–259.

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

Bochner, J. H., Samar, V. J., Hauser, P. C., Garrison, W. M., Searls, J. M., & Sanders, C. A. (2016). Validity of the American Sign Language Discrimination Test. *Language Testing, 33*(4), 473–495.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101.

Bridgeman, B., Cho, Y., & DiPietro, S. (2016). Predicting grades from an English language assessment: The importance of peeling the onion. *Language Testing, 33*(3), 307–318.

Brooks, L., & Swain, M. (2014). Contextualizing performances: Comparing performances during TOEFL iBTTM and real-life academic speaking activities. *Language Assessment Quarterly, 11*(4), 353–373.

Butler, Y. G., & Zeng, W. (2014). Young foreign language learners' interactions during task-based paired assessments. *Language Assessment Quarterly, 11*(1), 45–75.

Cai, H. (2013). Partial dictation as a measure of EFL listening proficiency: Evidence from confirmatory factor analysis. *Language Testing, 30*(2), 177–199.

Cai, H. (2015). Weight-based classification of raters and rater cognition in an EFL speaking test. *Language Assessment Quarterly, 12*(3), 262–282.

Canale, M., & Swain, M. (1981). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1–47.

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

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing, 32*(3), 385–405.

Chapman, E. (2003). Alternative approaches to assessing student engagement rates. *Practical Assessment, Research & Evaluation, 8*(13), 1–7.

Choi, I. (2017). Empirical profiles of academic oral English proficiency from an international teaching assistant screening test. *Language Testing, 34*(1), 49–82.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117–135.

Denies, K., & Janssen, R. (2016). Country and gender differences in the functioning of CEFR-based can-do statements as a tool for self-assessing English proficiency. *Language Assessment Quarterly, 13*(3), 251–276.

Dulock, H. L. (1993). Research design: Descriptive research. *Journal of Pediatric Oncology Nursing, 10*(4), 154–157.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly, 9*(3), 270–292.

Eckes, T. (2014). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing, 31*(1), 39–61.

Eckes, T. (2017). Setting cut scores on an EFL placement test using the prototype group method: A receiver operating characteristic (ROC) analysis. *Language Testing, 34*(3), 383–411.

Farnsworth, T. L. (2013). An investigation into the validity of the TOEFL iBT speaking test for international teaching assistant certification. *Language Assessment Quarterly, 10*(3), 274–291.

Fidalgo, A. M., Alavi, S. M., & Amirian, S. M. R. (2014). Strategies for testing statistical and practical significance in detecting DIF with logistic regression models. *Language Testing, 31*(4), 433–451.

Goodwin, A. P., Huggins, A. C., Carlo, M., Malabonga, V., Kenyon, D., Louguit, M., & August, D. (2012). Development and validation of extract the base: an English derivational morphology test for third through fifth grade monolingual students and Spanish-speaking English language learners. *Language Testing, 29*(2), 265–289.

Granfeldt, J., & Ågren, M. (2014). SLA developmental stages and teachers' assessment of written French: Exploring Direkt Profil as a diagnostic assessment tool. *Language Testing, 31*(3), 285–305.

Green, A., & Hawkey, R. (2012). Re-fitting for a different purpose: A case study of item writer practices in adapting source texts for a test of academic reading. *Language Testing, 29*(1), 109–129.

Han, C. (2016). Investigating score dependability in English/Chinese interpreter certification performance testing: A generalizability theory approach. *Language Assessment Quarterly, 13*(3), 186–201.

Harding, L. (2014). Communicative language testing: Current issues and future research. *Language Assessment Quarterly, 11*(2), 186-197.

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing, 32*(3), 317–336.

Haug, T. (2012). Methodological and theoretical issues in the adaptation of sign language tests: An example from the adaptation of a test to German Sign Language. *Language Testing, 29*(2), 181–201.

Hirai, A., & Koizumi, R. (2013). Validation of empirically derived rating scales for a story retelling speaking test. *Language Assessment Quarterly, 10*(4), 398–422.

Hoang, G. T. L., & Kunnan, A. J. (2016). Automated Essay Evaluation for English Language Learners: A Case Study of MY Access. *Language Assessment Quarterly, 13*(4), 359–376.

Hsieh, M. (2013a). An application of Multifaceted Rasch measurement in the Yes/No Angoff standard setting procedure. *Language Testing, 30*(4), 491–512.

Hsieh, M. (2013b). Comparing yes/no Angoff and Bookmark standard setting methods in the context of English assessment. *Language Assessment Quarterly, 10*(3), 331–350.

Hsu, T. H. L. (2016). Removing bias towards World Englishes: The development of a Rater Attitude Instrument using Indian English as a stimulus. *Language Testing, 33*(3), 367–389.

Huang, B., Alegre, A., & Eisenberg, A. (2016). A cross-linguistic investigation of the effect of raters' accent familiarity on speaking assessment. *Language Assessment Quarterly, 13*(1), 25–41.

Huang, F. L., & Konold, T. R. (2014). A latent variable investigation of the Phonological Awareness Literacy Screening-Kindergarten assessment: Construct identification and multigroup comparisons between Spanish-speaking English-language learners (ELLs) and non-ELL students. *Language Testing*, *31*(2), 205–221.

Ilc, G., & Stopar, A. (2015). Validating the Slovenian national alignment to CEFR: The case of the B2 reading comprehension examination in English. *Language Testing*, *32*(4), 443–462.

Jin, T., & Mak, B. (2013). Distinguishing features in scoring L2 Chinese speaking performance: How do they work?. *Language Testing*, *30*(1), 23–47.

Jin, T., Mak, B., & Zhou, P. (2012). Confidence scoring of speaking performance: How does fuzziness become exact?. *Language Testing*, *29*(1), 43–65.

Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly*, *9*(3), 249–269.

Katzenberger, I., & Meilijson, S. (2014). Hebrew language assessment measure for preschool children: A comparison between typically developing children and children with specific language impairment. *Language Testing*, *31*(1), 19–38.

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly*, *12*(3), 239–261.

Knoch, U., & Chapelle, C. A. (2017). Validation of rating processes within an argument-based framework. *Language Testing*, *34*, 1–23.

Kokhan, K. (2013). An argument against using standardized test scores for placement of international undergraduate students in English as a Second Language (ESL) courses. *Language Testing*, *30*(4), 467–489.

Koo, J., Becker, B. J., & Kim, Y. S. (2014). Examining differential item functioning trends for English language learners in a reading test: A meta-analytical approach. *Language Testing*, *31*(1), 89–109.

Kuiken, F., & Vedder, I. (2017). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing*, *34*(3), 321–336.

Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing*, *33*(3), 319–340.

Lado, R. (1961). *Language Testing*. New York: McGraw-Hill.

Lam, R. (2015). Language assessment training in Hong Kong: Implications for language assessment literacy. *Language Testing*, *32*(2), 169–197.

Lee, H., & Winke, P. (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.

Lee, S., & Winke, P. (2018). Young learners' response processes when taking computerized tasks for speaking assessment. *Language Testing*, *35*(2), 239-269.

Li, H., & Suen, H. K. (2013). Detecting native language group differences at the subskills level of reading: A differential skill functioning approach. *Language Testing*, *30*(2), 273–298.

Li, H., Hunter, C. V., & Lei, P. W. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing*, *33*(3), 391–409.

Lin, C. K., & Zhang, J. (2014). Investigating correspondence between language proficiency standards and academic content standards: A generalizability theory study. *Language Testing, 31*(4), 413–431.

Ling, G. (2017). Is writing performance related to keyboard type? An investigation from examinees' perspectives on the TOEFL iBT. *Language Assessment Quarterly, 14*(1), 36–53.

Mann, W., Roy, P., & Morgan, G. (2016). Adaptation of a vocabulary test from British Sign Language to American Sign Language. *Language Testing, 33*(1), 3–22.

Murray, J. C., Riazi, A. M., & Cross, J. L. (2012). Test candidates' attitudes and their relationship to demographic and experiential variables: The case of overseas trained teachers in NSW, Australia. *Language Testing, 29*(4), 577–595.

Nakatsuhara, F., Inoue, C., Berry, V., & 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.

Pan, M., & Qian, D. D. (2017). Embedding Corpora into the Content Validation of the Grammar Test of the National Matriculation English Test (NMET) in China. *Language Assessment Quarterly, 14*(2), 120–139.

Papageorgiou, S., & Cho, Y. (2014). An investigation of the use of TOEFL® Junior™ Standard scores for ESL placement decisions in secondary education. *Language Testing, 31*(2), 223–239.

Pill, J., & McNamara, T. (2016). How much is enough? Involving occupational experts in setting standards on a specific-purpose language test for health professionals. *Language Testing, 33*(2), 217–234.

Saida, C. (2017). Creating a Common Scale by Post-Hoc IRT Equating to Investigate the Effects of the New National Educational Policy in Japan. *Language Assessment Quarterly, 14*(3), 257–273.

Sanchez, S. V., Rodriguez, B. J., Soto-Huerta, M. E., Villarreal, F. C., Guerra, N. S., & Flores, B. B. (2013). A case for multidimensional bilingual assessment. *Language Assessment Quarterly, 10*(2), 160–177.

Sato, T. (2012). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing, 29*(2), 223–241.

Savignon, S. J. (1972). *Communicative competence: An experiment in foreign language teaching*. Philadelphia: Center for Curriculum Development.

Shaw, S., & Imam, H. (2013). Assessment of international students through the medium of English: Ensuring validity and fairness in content-based examinations. *Language Assessment Quarterly, 10*(4), 452–475.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*(1), 27–33.

Spolsky, B. (2008). Language assessment in historical and future perspective. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of language and education* (Second ed., Vol. 7: Language testing and assessment, pp. 445–454). New York: Springer Science.

Stansfield, C. W. (2008). Lecture: "Where we have been and where we should go". *Language Testing, 25*(3), 311–326.

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.