English Language Teaching; Vol. 12, No. 12; 2019
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

Language Testing in China: Past and Future

Xuelian Li[1]

[1] Faculty of English Language and Culture, Guangdong University of Fereign Studies, Guangzhou, China

Correspondence: Xuelian Li, Northern Baiyun Avenue, Baiyun District, Guangzhou, China.

Received: October 15, 2019 Accepted: November 13, 2019 Online Published: November 15, 2019

doi: 10.5539/elt.v12n12p67 URL: https://doi.org/10.5539/elt.v12n12p67

Abstract

Based on the articles written by mainland Chinese scholars published in the most influential Chinese and international journals, the present article analyzed the language testing research, compared the tendencies of seven categories between 2000-2009 and 2010-2019, and put forward future research directions by referring to international hot topics. Of all the seven categories of research topics, validity, performance test and China’s Standards of English Language Ability were three most popular themes, while classroom assessment, technology, rater/test taker differences and professionalization were much less popular. Except for research on performance test and technology, the other five aspects showed an increase in the second decade, with that of China’s Standards of English Language Ability rising the most dramatically. Referring to international research trends, the research predicted that validity, classroom assessment, China’s Standards of English Language Ability and professionalization, especially the ethics and social justice, might be the promising research topics for language testers.

Keywords: language testing, validity, professionalization, CSE, classroom assessment

1. Introduction

English tests have become essential for Chinese to get the opportunities of further education, job application and promotion. Chinese take part in big-scale high-stakes tests designed by Chinese researchers, i.e., College English Test (CET), Matriculation English Test for College Entrance (MET), Test for English Majors (TEM) and Public English Test System (PETS), and international organizations, i.e., TOEFL (Test of English as a Foreign Language), IELTS (Test of English as a Foreign Language) (Yu & Jin, 2016). Additionally, the relatively low-stakes assessment, i.e., classroom assessment, gained its popularity in recent years. In this sense, language testing in this article includes high-stakes and low-stakes as well.

With the increase of test takers and test types, the research on language testing extends to various aspects, especially since the advent of 21st century (Mao & Wu, 2015). It is, therefore, necessary to figure out the trends of relevant research and provide directions for future research based on the analysis of the published influential articles in Chinese and international journals.

Previous studies investigated tendencies of language testing in China in recent decades, but some of them (Hu, 2007; Wu, 2009) did the research ten years ago, so their research results do not fit the current situation. Some other articles (Jia, 2013; Jiang, 2018; Mao & Wu, 2015) are more up to date, but only Jiang’s (2018) work can be used as a reference for researchers, since dramatic changes happened in China in several years, due to the initiation of Chinese educational reforms in testing and assessment (State council, 2014).

Furthermore, all of the aforementioned articles were confined to articles published in Chinese journals, while many articles published in international top language testing journals were written by big names, who lead directions of language testing in China. Therefore, the above mentioned articles did not show the whole picture of research on Chinese language testing.

Besides the incomplete summary of Chinese language testing research, those articles could not make a reasonable and comprehensive comparison, analysis and summary, either, since they did not take international research into consideration. The present research, therefore, intends to investigate the current trends of language testing in China and provide the future directions on the basis of analyzing international studies.

The research set the time slot of 20 years, between 2000 and 2019, because Bachman (2000) summarized language testing research in the world before the year of 2000. As he stated, the first ten years, the 1980s, witnessed the expansion of language testing research on communicative language competence, test-taking strategies, bilingual proficiency and language learners' developmental sequence. The trends continued in the following decade, which saw further expansions in these five areas:

a) research methodology;

b) practical advances;

c) factors that affect performance on language tests;

d) authentic, or performance, assessments; and

e) concerns with the ethics of language testing and professionalizing the field (Bachman, 2000).

The seminal article also concluded with two future directions: the professionalization of the field and validation research. Professionalization of the field is "concerned with codes, contracts, professional training, professional ethical norms and standards, and the systematic attempt to illuminate the ethics of a profession and to elaborate its norms" (p. 19). It includes training of language testing professionals and developing relevant standards of practice and mechanisms.

Bachman (2000) thought validation had been and would continue to be a recurring research topic. The future validation research, however, can be conducted with a breadth of coverage, including factors/processes affecting test performance, application of different research tools, the political/ethical issues of test use, and construct interpretations from the consequences of test use. The validation should also focus on technology in language testing.

2. Classification and Research Questions

It has been twenty years since the publication of the pivotal paper. This article intended to find out how much research on professionalization of language testing and validation, which were the research directions raised by Bachman (2000), has been conducted by Chinese scholars and how the research has been progressed.

Moreover, the current research overviewed all the research conducted during twenty years, so there should be some more types besides the above two. In reality, the validation research by Bachman (2000) is an integrated concept, covering almost all the aspects in test development and use. Language testing research combines the fields of linguistics, language teaching and psychometrics, so it can be divided into an abundance of categories with various classifying methods. However, the classification will not be much practical when the studies are divided into too many themes. For example, Jiang (Jiang, 2018) collated 16 themes based on 11 years' language research, but they were integrated into six types when she tried to explain the tendencies. Moreover, some of the integration is far-fetched and weird. For instance, she merged three completely different themes -internet/web-based test, construct and CSE - into one aspect. In view of this, the current study narrowed down the list to several highlighted themes, i.e., the themes much frequently investigated. Therefore, the researcher integrated two recently developed classification schemes, Plakans (2018) and Aryadoust and Fox (2016). The former compared the main tendencies of language testing research of Year 2007 and Year 2017 in four broad themes – validity, performance test, classroom assessment and technology. The latter, classified the research from the regions of Middle East and Pacific Rim into five categories: technology, test takers & raters, diagnostic assessment, construct, and impact & washback.

Obviously, the theme of technology appeared repeatedly in those two classification schemes, and construct and impact and washback were merged into the theme of validity, since they are generally the subdivision of validity (Messick, 1989). Furthermore, the theme of diagnostic assessment were sorted together with classroom assessment. Consequently, the integration made five categories.

Professionalisation is the future direction designated by Bachman (2000), so it should be in the list. Moreover, China's Standards of English Language Ability (CSE hereafter, cf. 4.2 & 4.3 for more information) is another rising research field, since the nation-wide research and creation of the English standard brought about some and will bring out more radical changes in English teaching, learning and testing in China (Liu, 2017). Henceforth, the list will include seven categories - validity, performance test, classroom assessment, technology, rater/test taker differences, CSE, professionalization.

Based on the above seven categories, the research aimed to answer the following three research questions:

1) How much language testing research has been conducted on these seven categories?

2) What are the research tendencies of these seven categories?

3) What are the future directions for language testing researchers?

The first two research questions were answered by surveying literature, counting the frequency and finding out the tendencies of those seven categories. The status quo of language testing in China, together with points of view from famous specialists, were integrated, so as to offer future research directions.

3. Coding Procedures

The articles, published from 2000 to 2019, were mainly chosen from three sources: ILTA (International Language Testing Association) language testing bibliography, five top international journals of language testing (*Language Testing, Language Assessment Quarterly, Language Testing in Asia, Papers in Language Testing and Assessment and Assessing Writing*), and three Chinese top journals of linguistics (*Modern Foreign Languages, Foreign Language Teaching and Learning, Foreign Language Field*). The most seminal and rigorous works by Chinese scholars and specialists on language testing can be found on the above three sources.

The articles were selected based on five criteria. That is, they are peer-reviewed journal articles by Chinese mainland researchers conducted in mainland China from 2000 to 2019. To be specific, the criteria are listed as follows:

1) The time span of the articles should be from 2000 to 2019;

2) The articles should be empirical articles published on journals;

3) They do not include book reviews, book sections, or unpublished doctoral dissertations;

4) The affiliation of first authors should be in mainland China;

5) The research data should be collected in mainland China.

On the basis of the criterion, 109 articles were found from ILTA (International Language Testing Association) website, 24 from the five top international journals, 196 articles from three Chinese journals. Altogether 329 articles were sorted together. The next step was to delete the repeated articles in the three lists, because some articles from ILTA language testing bibliography reappeared in the international and Chinese journal article list. They were dropped out one by one. In the end, 269 articles remained for analysis.

Then, each category was numbered consecutively, so the numbering was like this: 1-validity, 2-performance test, 3-classroom assessment, 4-technology, 5-rater/test taker differences, 6-China's Standards of English Language Ability, 7-professionalization. Each category was further divided into two parts to compare research trends in the two decades to answer Research Question Two. The first part was between 2000 and 2009, and the second part was between 2010 and 2019. For example, a validation article published in 2002 was coded as 1-1, while that in 2012 was coded as 1-2.

Priority was given to the category with a bigger number if one article covers two topics. Each of the seven topics was represented by a number from one to seven (cf. Figure 1), and the articles with two topics was classified as the category of bigger number. For instance, the article entitled *Computer literacy and the construct validity of a high-stakes computer-based writing assessment* (Jin & Yan, 2017) covers the topics of technology and validity, but it was sorted into technology, because four represents technology and one represents validity (cf. Figure 1), and because four is bigger than one.

4. Results and Discussion

4.1 Research on Seven Categories

Of the 269 articles, 217 were successfully collated into the seven types. The other 52 articles were unable to form any groups, in view of the small number, so they were put aside. The 217 articles spread unevenly, as can be seen obviously in Table 1. What stand out in the figure are studies on validity, performance test and China's Standards of English Language Ability. Validity research reached the peak of all the seven categories, though some of the interdisciplinary studies were all coded into subsequent groups according to the coding method.

This big amount of validity research, with 49 publications, supported Bachman's (2000) summary and prediction that validity was and would still be research focus. Validity is "the most important quality"(Bachman, 1990: p. 289) in developing, interpreting and using language tests. Investigating validity of language tests, therefore, is a necessary and long-lasting process. What's more, validity is inclusive, with aspects ranging from content validity, construct validity, face validity, washback and so on (Messick, 1989).

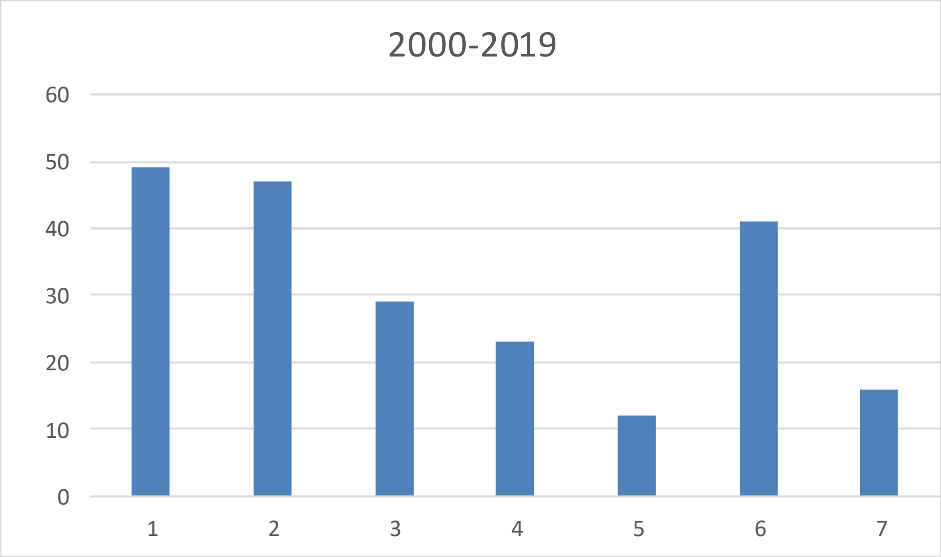

Figure 1. The seven categories of language testing research

Note: 1-validity, 2-performance test, 3-classroom assessment, 4-technology, 5-rater/test taker differences, 6-China's Standards of English Language Ability, 7-professionalization.

Since the beginning of the new century, validity research in China was very extensive. Much work was conducted on washback research (Jin, 2000; Li, 2005; Qi, 2004a, 2004b, 2005, 2007, 2012; Zhan & Andrews, 2014; Zou & Xu, 2017). Others were about construct validity (Jin & Yan, 2017; Liu & Yang, 2001), content validity (Pan & Qian, 2017), predictive validity (Fan, 2016) and the validation of self-developed instruments (Liu, 2007; Zhao, 2012).

Performance test is the second biggest test type, with only two articles fewer than that of validity. It is the test where "test takers' performance is expected to replicate their language performance in non-test situations" (Bachman, 1990: p. 77). Influenced by the concept of communicative language ability, increasing researchers started to design authentic tasks, so research on performance test attracted attention during the last two decades, covering writing (Chen, 2010), speaking (He & Zhang, 2008) and translation (Mu, 2006).

The least investigated research theme, of the seven themes, is research on rater/test taker differences. In language testing, test takers' scores are affected by raters. Research found raters' cognitive and metacognitive strategies affected rating accuracy (Zhang, 2016). Meanwhile, test methods and other physical factors may affect test results, that's why test takers, with different language proficiency, were affected by task types (Zhang & Zhou, 2014).

To sum up, the seven types of research varied in terms of total number. The top three widely researched categories are on validity, performance test and CSE, with over 40 publications for each type. And the least category is rater/ test taker differences, with only 12 articles published in total.

4.2 Tendencies of Seven Categories

The comparison results of articles during the two stages, 2000-2009 and 2010-2019, show the tendencies of seven categories (cf. Figure 2). Obviously, five categories rose up during the second decade, except for performance test and technology. The most dramatic increase is the research on CSE. Fang, Yang, and Zhu (2008) discussed the necessity, principles and methods of creating a national language scale as a systematic criterion to guide language teaching and learning. With the message sent in the article, and the policy requirement (State Council, 2014), China commenced the historical large-scale (with the participation of more than 300 scholars and the data collected from 28 out of 34 provincial regions in China) language framework development (Liu & Peng, 2017), to pave the way for establishing a national foreign language assessment system. It is hoped that the new language assessment system will replace a multitude of repeatedly unnecessary English tests in China. After three years of hard work, China's Standards of English Language Ability was successfully created, with both Chinese and English versions released (Liu et al., 2018). Along with the progress of the project, there has been a steep rise in the amount of research since 2015. Studies ranged from the framework exploration of overall scales (Liu, 2015; Liu & Han, 2018; Zhu, 2016) to the sub-scales i.e., the speaking scale (Jin & Jie, 2017), reading scale (Zeng, 2017), vocabulary scale (Zhao, Wang, Coniam, & Xie, 2017), writing strategy scale (Deng & Deng, 2017) and the scale of translation and interpretation (Bai, Hong, & Yan, 2018; Wang, Xu, & Mu, 2018). Afterwards, researchers from international tests, i.e., IELTS, Aptis (developed by British Council) and TOEFL, carried out the alignment research with counterparts from CSE group. The epochal work then generated subsequent validation research CSE (Cai, 2019; He, 2019).

The research on classroom assessment, rater/test taker differences and professionalization also increased sharply, with research on classroom assessment and professionalization doubling and rater/test taker differences tripling in the latest decade. Meanwhile, more work was carried out on validity, with a growth of 11 articles between 2010 and 2019, compared to that in previous ten years.

On the contrary, research on performance test and technology fell down in the later decade. The former decreased slightly to 22 from the original 25. The latter had a salient decrease, from 15 in the first period to eight in the second one. The possible reason for falling technology research in language testing may be the periodical interest in computer/web use. The first period saw the surging application of computer in daily life and office work in China, so the advent of computer in language testing was a novelty. In this vein, much research introduced the use of automatic scoring system (Ge & Chen, 2007; Wang & Wen, 2009). After 10 years' widespread public interest in computer, computer-based language testing is not new to learners and language testers. However, studies from the new perspective, such as learning-based automatic scoring system (Li & Liu, 2013; 2016) and computer literacy in internet-based tests (Jin & Yan, 2017), will be attractive in the future.

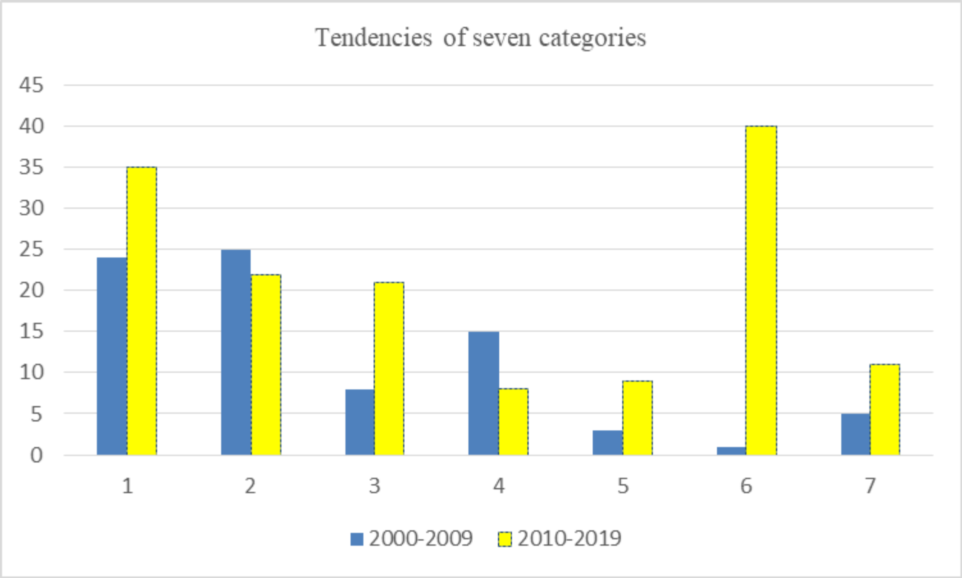

Figure 2. The tendencies of language testing research

Note: 1-validity, 2-performance test, 3-classroom assessment, 4-technology, 5-rater/test taker differences, 6-China's Standards of English Language Ability, 7-professionalization.

To sum up, five out of seven themes, in the second decade, showed a rising tendency, but research on performance test and technology decreased, compared to the studies in the previous decade. On the whole, there has been a radical increase in the number of language testing research in China since 2010, with an increase of 65 published articles in all.

4.3 The Future Research Directions

As mentioned above, validity was, and will still be, research focus of language testing in China, in view of the necessity and importance of the quality accounting for language test development, interpretation and use.

The research of professionalization, the future direction offered by Bachman in 2000, was not extensively studied in China, though some work has been conducted to explore assessment literacy (Xu & Brown, 2017; Xu & Brown, 2016) and teachers' classroom practice (Wang, 2017). The standard for language teachers' professional development is in need (Fulcher, 2012). More urgently needed is field research on how to help teachers internalize the theoretical knowledge into their instruction (Andersson & Palm, 2018), especially for them to implement classroom assessment appropriately (Brown, 2017) in the complex and conflicting educational context (Qi, 2005; Vandeyar, 2005).

Professionalization also concerns research on fairness in language testing and social value/impact of language testing. Any high-stakes tests influence different stakeholders, test takers, teachers, school administrators and educational administrators. China usually has a large population of test takers, who own different physical characteristics and come from various regions, so the problem of fairness should not be ignored by language testers, and language testers should help make reasonable language testing policy (Elder & Harding, 2008; Ross, 2008).

In the future, near or far, research on technology of language testing will continue to be the zest for innovative language testers. Research can be done to compare differences between computer-based tests and traditional paper-and-pencil tests and fairness of scoring (Yu & Zhang, 2017). However, focus should be put on security issues of computer-based and web-based tests (Chapelle & Voss, 2016), the expansion of passage scoring abilities, and integration of automatic assessment with formative assessment (Liu, 2013).

The roaring increase of research on CSE also signifies that CSE is a promising field, since corresponding language proficiency tests, covering Levels Four to Eight, will be released. The newly developed tests will provide ample empirical data for research, and the validation research will be used to refine the tests. Furthermore, CSE includes eight self-assessment scales (Liu et al., 2018), designed for teachers and students' formative use, more field work can be conducted on how the scales are applied in English teaching and learning. In addition, practical and valuable suggestions are welcomed to refine CSE.

5. Conclusion

The present article analyzed the language testing research conducted by Chinese researcher during the last 20 years, compared the tendencies of seven categories between the two decades, and put forward the future research directions based on international language testing research. Validity, the everlasting field, will accompany the evolvement of language testing. Research on classroom assessment and CSE will continue to rise not only because of the orientation of national policy, but also because of the international trend of language testing focus on practical problems (Bachman, 2000). The research on professionalism, especially ethics and social fairness in language testing, needs to expand in view of the relatively scarce research in China and its significance in the world. The conference themes of both recent LTRC (Language Testing Research Colloquium) conferences, 2019 and 2020, concern the discussion of social justice. Bachman (2000) also emphasized that fairness was "at the heart of how we define ourselves as professionals" (p. 25). As a professional practice, language testing should primarily involve the issues of fairness and ethics. In addition, language testers should also set/hold professional standards, and address the accountability to society and by society in developing and using language tests.

Acknowledgments

The author would like to express her thankfulness to The International Cultivation Project of Excellent Young Scholars (Group Three) by Guangdong Provincial Educational Bureau, Postgraduate Research & Innovation Project of Guangdong University of Foreign Studies (18GWCXXM-41), and Fund for Assessment Research in Foreign Language Education (FAR201702009) for funding this research. My special thanks go to Jianda Liu, who encouraged me to read widely and helped with the preliminary data collection.

References

Andersson, C., & Palm, T. (2018). Reasons for teachers' successful development of a formative assessment practice through professional development – a motivation perspective. *Assessment in Education: Principles, Policy & Practice, 25*(6), 576-597. https://doi.org/10.1080/0969594X.2018.1430685

Aryadoust, V., & Fox, J. (2016). *Trends in language assessment research and practice: The view from the Middle East and the Pacific Rim*. Cambridge Scholars Publishing.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford University Press

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.

Bai, L., Hong, L., & Yan, M. (2018). The construct and principles of translating abililities of CSE. *Modern Foreign Languages, 41*(01), 101-110+147.

Brown, J. D. (2017). Forty years of doing second language testing, curriculum, and research: So what? *Language Teaching, 50*(2), 276-289. https://doi.org/10.1017/S0261444816000471

Cai, H. (2019). The validity of aligning productive language tests with CSE: Generalizability and consistency. *Modern Foreign Languages, 42*(05), 709-721.

Chapelle, C. A., & Voss, E. (2016). Utilizing technology in language assessment. *Language Testing and Assessment*, 1-13. https://doi.org/10.1007/978-3-319-02326-7_10-1

Chen, H. (2010). The types and characteristics of measuring indexes of English writing. *Modern Foreign Languages, 33*(1), 72-80.

Deng, J., & Deng. H. (2017). The framework of writing strategy scale in CSE. *Foreign Language Field, 182*(02), 29-36.

Elder, C., & Harding, L. (2008). Language testing and English as an international language. *Australian Review of Applied Linguistics, 31*(3), 34.31-34.11. https://doi.org/10.2104/aral0834

Fan, J. J. (2016). The construct and predictive validity of a self-assessment scale. P*apers in Language Testing and Assessment, 5*(2), 69-100.

Fang, X., Yang, H., & Zhu, Z. (2008). The principles and methods of creating a national standard of language ability. *Modern Foreign Languages, 31*(4), 380-387.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113-132. https://doi.org/10.1080/15434303.2011.642041

Ge, S., & Chen, X. (2007). The exploration of Chinese EFL learners' automatic scoring of compositions. *Foreign Language Field, 172*(05), 43-50.

He, L., & Zhang, J. (2008). The reliability of oral tests in CET 4 & 6: The approach of Rasch Model. *Modern Foreign Languages, 31*(04), 388-398+437.

He, L. (2019). The validation of aligning language tests with CSE. *Modern Foreign Languages, 42*(5), 660-671.

Hu, H. (2007). The review of English tests in China. *Journal of Hunan College of Arts and Science (Social Science), 32*(01), 137-139.

Jia, W. (2013). The metrological analysis of language testing articles in China in the recent decade. *Journal of Changchun University of Science and Technology (Social Science), 26*(12), 134-136+156.

Jiang, J. (2018). The empirical language testing research: Past and future. *Foreign Language Field, 183*(02), 40-48.

Jin, Y. (2000). Washback of speaking tests in CET 4 & 6. *Foreign Language Field, 165*(04), 56-61.

Jin, Y., & Jie, W. (2017). The principles and methods of creating speaking scale in CSE. *Foreign Language Field, 182*(02), 10-19.

Jin, Y., & Yan, M. (2017). Computer Literacy and the Construct Validity of a High-Stakes Computer-Based Writing Assessment. *Language Assessment Quarterly, 14*(2), 101-119. https://doi.org/10.1080/15434303.2016.1261293

Li, S. (2005). Washback in language testing and test design. *Foreign Language Field, 170*(01), 71-75.

Li, X., & Li, J. (2013). Ensemble learning based essay automated scoring algorithm for Chinese English learners. *Journal Of Chinese Information Processing, 27*(5), 100-106.

Li, X., & Liu, J. (2016). Automatic Essay Scoring Based on Coh-Metrix Feature Selection for Chinese English Learners. Paper presented at the International Symposium on Emerging Technologies for Education. https://doi.org/10.1007/978-3-319-52836-6_40

Liu, J. (2007). Developing a pragmatics test for Chinese EFL learners. *Language Testing, 24*(3), 391-415. https://doi.org/10.1177/0265532207077206

Liu, J. (2013). Modern Educational Technology and Language Testing. *Technology Enhanced Foreign Language Education, 152*(04), 46-51.

Liu, J. (2015). The foundimental methods of creating CSE. *China Examination, 2015*(01), 7-11+15.

Liu, J. (2017). CSE and English learning. *Foreign Languages in China, 14*(06), 4-11.

Liu, J., & Han. B. (2018). The theoretical foundations of creating practical CSE. *Modern Foreign Languages, 41*(01), 78-90+146.

Liu, J., & Peng, C. (2017). Creating a scientific CSE. *Foreign Language Field, 182*(2), 2-9.

Liu, J., & Yang, M. (2001). What does proofreading measuring? *Modern Foreign Languages, 24*(2), 170-180.

Liu, J., et al. (2018). *China's Standards of English Language Ability*. (GF 0018-2018). Beijing. Retrieved from http://www.neea.edu.cn/res/Home/1908/0c96023675649ac8775ff3422f91a91d.pdf

Mao, Y., & Wu, Q. (2015). A review of language testing in China during the recent 20 years. *Foreign Language Education*, 167-180.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: Macmillan.

Mu, L. (2006). Testing and scoring translation. *Foreign Language Teaching and Research, 38*(6), 466-471.

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. https://doi.org/10.1080/15434303.2017.1303703

Plakans, L. (2018). Then and Now: Themes in Language Assessment Research. *Language Education & Assessment, 1*(1), 3-8. https://doi.org/10.29140/lea.v1n1.35

Qi, L. (2004a). Has a high-stakes test produced the intended changes? *Washback in language testing: Research contexts and methods*, 171-190.

Qi, L. (2004b). A study on the washback of the NMET. *Foreign Language Teaching and Research, 36*(05), 357-363+401.

Qi. L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing, 22*(2), 142-173. https://doi.org/10.1191/0265532205lt300oa

Qi, L. (2007). Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China. *Assessment in Education, 14*(1), 51-74. https://doi.org/10.1080/09695940701272856

Qi, L. (2012). Recent studies on washback of language testing and its future. *Modern Foreign Languages, 35*(02), 202-208+220.

Ross, S. J. (2008). Language testing in Asia: Evolution, innovation, and policy challenges. *Language Testing, 25*(1), 5-13. https://doi.org/10.1177/0265532207083741

State Council of People's Republic of China. (2014). The plans on implementing testing and recruiting reforms (State Council). (000014349/2014-00102).

Vandeyar, S. (2005). Conflicting demands: Assessment practices in three South African primary schools undergoing desegregation. *Curriculum Inquiry, 35*(4), 461-481. https://doi.org/10.1111/j.1467-873X.2005.00337.x

Wang, J., & Wen, Q. (2009). The construction of computer-assisted scoring model in big-scale tests. *Modern Foreign Languages, 32*(4), 415-420.

Wang, W., Xu, Y., & Mu, L. (2018). The interpretation scale in CSE. *Modern Foreign Languages, 41*(01), 111-121+147.

Wang, X. (2017). A Chinese EFL Teacher's Classroom Assessment Practices. *Language Assessment Quarterly, 14*(4), 312-327. https://doi.org/10.1080/15434303.2017.1393819

Wu, D. (2009). The review on Chinese language testing in recent ten years. *Journal of Dalian University of Technology (Social Science), 30*(04), 119-123.

Xu, Y., & Brown, G. T. (2017). University English teacher assessment literacy: A survey-test report from China. *Papers in Language Testing and Assessment, 6*(1), 133-158.

Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education, 58*, 149-162. https://doi.org/10.1016/j.tate.2016.05.010

Yu, G., & Jin, Y. (2016). *Assessing Chinese Learners of English: The Language Constructs, Consequences and Conundrums—An Introduction Assessing Chinese Learners of English* (pp. 1-16): Springer. https://doi.org/10.1057/9781137449788_1

Yu, G., & Zhang, J. (2017). Computer-based english language testing in China: Present and future. *Language Assessment Quarterly, 14*(2), 177-188. https://doi.org/10.1080/15434303.2017.1303704

Zeng, Y. (2017). The principles and methods of creating reading scale of CSE. *Foreign Language Field, 182*(05), 2-11.

Zhan, Y., & Andrews, S. (2014). Washback effects from a high-stakes examination on out-of-class English learning: Insights from possible self theories. *Assessment in Education: Principles, Policy & Practice, 21*(1), 71-89. https://doi.org/10.1080/0969594X.2012.757546

Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assessing Writing, 27*, 37-53. https://doi.org/10.1016/j.asw.2015.11.001

Zhang, X., & Zhou, Y. (2014). The effect of task types on Chinese EFL students' writing performance. *Modern Foreign Languages, 37*(4), 548-558.

Zhao, C. G. (2012). Measuring authorial voice strength in L2 argumentative writing: The development and validation of an analytic rubric. *Language Testing, 30*(2), 201-230. https://doi.org/10.1177/0265532212456965

Zhao, W., Wang, B., Coniam, D., & Xie, B. (2017). Calibrating the CEFR against the China Standards of English for College English vocabulary education in China. *Language Testing in Asia, 7*(1). https://doi.org/10.1186/s40468-017-0036-1

Zou, S., & Xu, Q. (2017). A Washback Study of the Test for English Majors for Grade Eight (TEM8) in China—From the Perspective of University Program Administrators. *Language Assessment Quarterly, 1*4(2), 140-159. https://doi.org/10.1080/15434303.2016.1235170

Zhu, Z. (2016). The validity framework for CSE. *China Examination, 2016*(08), 3-13.

Copyrights