The Literature Review of Writing Test in Past Decade in China

Xinyu Guo[1], Xin Dong[2],*

[1]*College of Foreign Language, Northeast Forestry University, Harbin, China*
[2]*College of Foreign Language, Northeast Forestry University, Harbin, China*
*Corresponding author

Abstract: *Writing reflects not only learners' ability to use language, but also their character as a person. The study of writing test has important implications for both Second Language Acquisition and teaching process. This paper reviews the research findings on English writing test in China over the past decade, including assessment, validity and backwash effect. The results demonstrate that researches on assessment of writing test includes scoring dimensions, scoring approaches, development of scoring scales and so on. Moreover, in terms of the validity, cognitive validity, content validity, and context validity have all been researched. However, compared with the validity of different task types of writing test, there are relatively few studies on these three dimensions. However, structure validity has been rarely mentioned in the validity studies of the last decade. As for backwash effect, both positive and negative effect have been studied, but there are still no practical measures to effectively attenuate or circumvent negative backwash effect. In a word, the researches related to writing test in China have become more and more systematic in the last decade.*

Keywords: *Writing test, Assessment. Validity, Backwash effect*

1. Introduction

Writing test is an essential part of English teaching and Second Language Acquisition. Proper use of writing test can effectively assess the effectiveness of writing teaching and students' writing competence, thus promoting students' English learning and improving teachers' teaching quality. The research focuses on a review of literature on writing test on the CNKI from 2013 to 2022. A search on CNKI using the keyword "writing test" reveals 220 academic journals, 305 dissertations, and 2 conference papers related to writing tests. The relevant researches focuses on backwash effect, validity, teaching, assessment and different types of writing. Among all the researches, there are 56 academic journals closely related to writing tests, including 16 high level publications from CSSCI and PKU.

The paper centres around three research themes that account for a relatively large portion of the research, assessment, validity, and backwash effect. Firstly, assessment of writing test mainly involves specific scoring approach, scoring dimensions, development of scoring criteria, and design of scoring scales. All of these factors have a relatively large impact on whether a writing test can fairly and reasonably measure a student's writing ability and level of knowledge. Secondly, validity is the extent to which empirical evidence and theoretical rationale support the adequacy and appropriateness of inferences made or actions taken based on test scores or other assessments (Messick, 1989). Therefore, validity is also a popular topic in writing test research, including content validity, context validity, and so on. Another focus of research is the backwash effect of writing tests, and it refers to the influence that testing exerts on teaching and learning (Hughes, 1989). The research of backwash effect is useful for facilitating the teaching and learning process in second language acquisition[1-4].

2. The Focus of Writing Test Research

In past decade, the research on writing test in China mainly involves assessment, validity and backwash effect.

2.1. The Assessment

Over the past decade, a major area of research in writing tests has focused on scoring, primarily

involving scoring methods, scoring dimensions, scoring criteria, and scoring scales. The scoring of second language writing test mainly focuses on the scoring dimensions of language use, discourse and content (Cumming et al, 2000). Before 2013, however, there were few systematic studies on the assessment methods of content quality in writing tests in China. Based on this research gap, Liu li (2013) explored a new rating method——the Hierarchical Decision-Tree Approach for assessing content quality. In the empirical study, the new approach was compared with traditional holistic scoring approach by two experienced raters assessing 81 writing scripts of English major students. The study states that the proposed approach enjoy relatively higher inter-rater reliability and test-retest reliability compared with those of traditional scoring. Nevertheless, the Hierarchical Decision-Tree Approach only applies to argumentation. However, the genre, including argumentation, is a key characteristic of the text, whose role in writing test is not well investigated in the testing field. Shao Jian (2021) analyzed the scoring validity of a self-chosen-task writing test by means of MFRM, in which the facets of students, raters, tasks and scale are probed into. The research reveal that the tasks are not significantly different and are equal in judging students with different language proficiency. The MFRM outputs indices support the self-consistency in the very writing test and proves its validity[5-9].

In addition to the assessment approach, the scoring criterion is also a research focus and difficulty in the writing testing. Because writing assessment is highly subjective, it is necessary to develop scientific, reasonable and maneuverable scoring criteria to ensure the reliability and validity of scoring (Becker, 2016). To make the scoring criterion as uniform as possible, Wu Xuefeng (2018) constructed a specific writing test rating scale, in which independent writing and integrative writing are combined, and validate the corresponding rating scale through applying the Multi-facet Rasch Model. However, only the scoring criteria of summary writing has been studied, and it is necessary to apply the scoring scale to other writing questions to verify its validity and reliability. Based on the limitations of previous studies, Shi Jianxiong (2022) constructed a scoring scale for diagnostic writing tests. Shi developed the rating scale based on four dimensions: language accuracy, language complexity, content rationality and writing normalization, and used the Multi-facet Rasch model to test whether the scale can carry out effective writing evaluation. The weakness of this study is the small sample size and study size, thus the practicability of this scoring scale needs to be further studied and confirmed. Zou Shaoyan (2022) explored the scoring criteria applicable to the CET-4 writing test, namely, lexical range, syntactic complexity, linguistic accuracy, orthographic control, cohesion and coherence, rhetorical organization, linguistic appropriacy, content and idea, fluency and task fulfillment. The research result lays a foundation for the further construction of the sub-item scoring scale of CET-4 writing test. However, its validity has not been widely verified due to its small sample. Therefore, whether the scoring criteria is practical and reliable remains to be discussed. In addition, as for English for Specific Purpose, Jia Wenfeng (2018) studied the medical English writing scoring scale. The three dimensions of the scoring scale are language performance proficiency, understanding and mastery of the patient's condition and the structure and logic of case record. Although the validity and reliability of the scale have yet to be verified, it can draw more attention to this type of test, so that scholars can develop and study the test tasks and scale forms with industry-specific features to better serve language teaching and writing test.

The reason is the scoring deviation caused by the subjectivity of the scorers why it is extremely difficult to determine the scoring criteria and scale for writing tests in the testing field. Yang Zhiqiang (2018) studied the scoring deviation of adult English writing test with the Multi-facet Rasch Model for bachelor degree. The results reveals significant differences in rater's leniency, but these differences does not affect the overall rating validity and the self-consistency of rater ratings is good, with no significant central tendency or randomness. However, there is still very few scorers with poor internal consistency and high use of low scoring bands, and a few scorers have deviation for scoring writing text of students.

Based on the above analysis, it can be concluded that the domestic scoring research of writing test in the past decade has mainly focused on the formulation of scoring criteria and the development of scoring scales. In addition, the harsh model is frequently used in writing tests to measure the practicality and validity of the scales. On the contrary, assessment approach and genres of writing with high scoring validity have been less studied in the testing field. In fact, the highly subjective nature of writing test scoring has forced scholars to grapple with how to make writing tests a fairer and more equitable test using a consistent and specific scale, on the other hand, this issue has been difficult to achieve compared to other test types[10-16].

2.2. The Validity

One of the relatively popular research directions is validity in the field of writing tests from 2013 to 2022, including context validity, content validity, cognitive validity, validity of scoring scale and the most

widely studied validity of different writing types. However, there is no research on structure validity in writing tests in the last decade on CNKI.

As a type of language test with high stakes and large-scale characteristics in China, the writing of CET-4 and CET-6 have been the focus of research in the field of writing test. In order to ensure that the test is an accurate measure of the student's English ability to achieve positive feedback on their learning, Chen Chen (2013) conducted an empirical study on the context validity of the CET-4 English writing test. Chen adopted a combination of quantitative and qualitative analysis to indicate that students performed well on the CET-4 writing test and were able to perform at a more realistic writing level. In addition, the context validity of the writing component of CET-4, in general, is acceptable. However, the research is a decade old, and further evidence is needed on the contextual validity of today's CET-4 writing test. In the same year, Xie Wanyi (2013) explored the content validity of writing test in PRETCO by analyzing writing pattern and mode and proposed some ways to improve the content validity of the PRETCO's writing test. However, these methods of improving content validity have not been validated and therefore the methods proposed by Xie have not been promoted and widely used in recent years.

Regarding cognitive validity, Zhang Peixin (2022) proposed that the cognitive validity of a writing test refers to the extent to which cognitive features triggered by the task are aligned with theoretical expectations. Using the argument-based validation approach, Zhang systematically argues for the cognitive validity of the TEM-8 writing test by testing three underlying assumptions in the TEM-8 writing test cognitive validity validation framework. The research demonstrates that test-takers' cognitive processes were in alignment with current cognitive processing theory pertaining to integrated writing, and reflective of specific cognitive operations that the TEM-8 integrated writing task required. Moreover, differences in cognitive processes were found to relate to proficiency levels. The test-takers' cognitive processes had significant effects on TEM-8 writing scores. Overall, the writing test of TEM-8 has a good cognitive validity. This research is one of the more recognized studies on the cognitive validity of writing tests in the last decade, and it is very informative and can be said to expand a research direction for the subsequent research on writing tests in China[17-22].

There have been fewer studies related to the validity of scoring scales in the last decade. Zou Shaoyan (2019) examined the validity of the holistic scoring scale adopted in the CET-4 writing assessment with questionnaire survey and interviews. The research indicates that the rating scale adopted by the CET-4 writing is generally satisfactory in terms of its clarity and practicality, however, there remains much room for improvement with respect to its completeness, feedback, and usefulness in rater training. This study informs the development of future scoring criteria and scales for CET-4 and CET-6.

In terms of validity of reading-to-write integrated test task, Wu Yue (2017) conducted a validity study on the EFL test task in terms of examinees, raters, and scoring rubrics by employing Multi-Facet Rasch Measurement model. The results of the study showed that the integrated task is relatively effective in differentiating the overall writing ability of the students. The following year, Zhang Haifeng (2018) also studied the validity of the reading-to-write integrated test task and concluded that this type of writing task is more helpful for students' writing skills, moreover, students trained in this way have improved their vocabulary, grammatical accuracy, sentence complexity, and creative perspectives. Both studies demonstrate the validity of reading-to-write integrated test task, which is why the types of writing tests in recent years have been dominated by integrated writing tasks. This not only tests students' ability of language use and writing skills, but also their reading comprehension and dialectical thinking skills. Another type of writing test that is more widely applied to examine a student's integrative competence is outline-free writing. Gong Wenwen (2014) examined the validity of this type of writing test in college English teaching and proposed that the outline-free writing test is a way with fairly high validity to test students' writing skills, and this kind of writing tasks can also improve students' writing competence. However, in the past decade, the validity of writing tasks has only focused on reading-to-write integrated test task and outline-free writing, with less attention paid to other types of writing tests. In fact, there are various types of second language writing questions, such as outline essays, diagram essays, proposition essays, first and last sentence essays, and epistolary essays, but these types of writing tests have rarely been mentioned in the last decade of research.

2.3. The Backwash Effect

From 2013 to 2022, compared with the assessment and the validity, the backwash effect was the most frequently studied in the field of English writing tests in China. The influence of testing on teaching and learning is known as the backwash effect (Alderson & Wall, 1993). Up to now, the academic community has conducted a more in-depth theoretical construction on the nature, dimensions, influencing factors, and mechanisms of the backwash effect. Although one purpose of writing tests is to test students' writing competence and level of knowledge, the purpose of testing students is also to provide focus and difficult points as well as directions for further teaching. As a result, the backwash effect is inevitably and

reasonably a research focus and relatively a hot topic in the field of writing test. Research on the backwash effect in the last decade has focused on CET-4, CET-6, IELTS, TOEFL, Senior-High school Entrance Examination and College Entrance Examination, all of which are high-stakes exams. It is generally accepted that the higher the stakes of a test, then the stronger its backwash effect on teaching and learning[23-25].

The research of the impact of various types of tests on the teaching and learning of English at the college level accounts for the largest proportion in the writing tests. Among plenty of types of test, CET-4 and CET-6 are most frequently studied because College English Test is an essential means to evaluate and improve college English teaching. Xiao Wei (2022) explored and compared the backwash mechanisms of CET-6, IELTS and TOEFL writing modules. Based on Xiao's results of the research, it can be concluded that, in general, achievement tests use influenced language learning and test-oriented preparation, mediated by test value. Moreover, instrumental exams can directly affect language learning. CET-6, IELTS, and TOEFL all differ significantly in their backwash effect on language learning, test design on value and test value on test-oriented preparation. The research corroborates the specificity of backwash in writing modules and the variations of backwash among different tests. Although this study sheds light on the validity of writing test, Xiaowei did not take into account other important factors that may influence the backwash effect, such as test anxiety, during the research process. Therefore the results of this study need to be further verified. In addition to Xiao's research, Feng Xingyu (2020) proposed that the CET-4 has a positive backwash effect on learning in general, mainly in terms of providing students with motivation and goals for learning English writing and improving their English writing skills and overall English proficiency. Zhang Fen (2017) analyzed the positive and negative backwash effect of the CET-4 writing test on classroom writing teaching, and proposed countermeasures and suggestions to weaken the negative backwash effect, such as reforming teaching methods and models to keep up with the times. However, Zhang does not propose some relatively specific, effective and practical methods. On the contrary, Tian Jin (2016) analyzed the factors affecting the writing of CET by studying the positive and negative backwash effect through tracking survey, and explored the complementary effect of product-focused approach and process-focused approach in college English writing teaching and learning. This research proposes some effective and practical teaching methods to weaken the negative backwash effect. For College Entrance Examination, Wang Ziwei (2020) and Xia Rui (2018) studied the backwash effect of writing test on the teaching and learning.

In response to the negative backwash effect that writing tests can have on teaching and learning, Zheng Hong (2013) and Xu Zongrui (2018) have done corresponding research. The negative backwash effect of writing tests on high-stakes exams, as indicated by Xu in his research, are manifested on following several levels: firstly, writing has not been paid enough attention to in teaching and learning, and secondly, test-taking skills have always been the theme of the teaching and learning. Zheng Hong made similar arguments and corresponding measures against the negative backwash effect of writing tests back in 2013. Zheng suggested that in order to weaken the negative backwash effect, it is necessary to increase the input of language materials and strengthen the teaching of discourse in the teaching process, on the other hand, it is also indispensable to develop students' critical thinking ability. In fact, writing test is a double-edged sword, and as research has shown, one of the most prominent negative backwash effect is that such test-oriented instruction tends to limit students' critical thinking ability. However, it is also undeniably difficult to completely circumvent the negative backwash effect[26-27].

3. Conclusions

From the above analysis, it can be concluded that the research outputs in writing tests in the last decade are increasingly systematic, including assessment, validity, backwash effect and so on. In terms of scoring, there is the Hierarchical Decision-Tree Approach that are relatively effective for scoring specifically argumentation, and increasingly objective and specific scoring criteria for different types of writing tasks. In addition, some researchers have developed scoring scales for different types of writing and have tested the validity of the scales. All of these research achievements are beneficial in promoting fair and objective assessment of writing tests. In terms of the validity of writing test, cognitive validity, content validity, and context validity have all been researched. However, compared with the validity of different task types of writing test, there are relatively few studies on these three dimensions. Moreover, structure validity has been rarely mentioned in the validity studies of the last decade. Comparatively speaking, the study of backwash effect accounts for a significant proportion of the research on the overall writing test field, and both positive and negative backwash effect have been studied. However, few researchers have proposed some practical, specific and effective measures on how to weaken the negative backwash effect on teaching and learning, rather than proposing ideas that are difficult to achieve.

In a word, writing test is steadily improving in China, as evidenced by increasingly objective and effective approach of assessment and reliable validity of writing test, and the fact that the existing writing

test largely have a positive backwash effect on teaching and learning. It is believed that the deficiencies and weaknesses of the existing research mentioned above will also be addressed in the future, thus contributing to a great extent to the flourishing development of the writing test field in China.

Acknowledgements

References

[1] Alderson J. and Wall D. Applied Linguistics [M]. London: Oxford University Press, (1993).
[2] Becker A. Student-Generated Scoring Rubrics: Examining their Formative Value for Improving ESL Students' Sriting Performance [J]. Assessing Writing, (2016)29: 15-24.
[3] Cumming A., Powers R., Santos T. and Taylor C. TOEFL 2000 Writing Framework: A Working Paper [J]. TOEFL Monograph 18, (2000).
[4] Hughes A. Testing for Language Teachers [M]. Cambridge: Cambridge University Press, (1989).
[5] Messick S. Validity and backwash in Language Testing [J]. ETS Research Report Series, (1996).
[6] Chen Chen. A Survey Study on the Context Validity of the Writing Test of CET-4[J]. Education and Teaching Research, (2013), 27:84-89.
[7] Xingyu Feng. An Empirical Study on the Backwash Effect of Writing Test on College Students' Writing Learning [J]. Journal of Kaifeng Vocational College of Culture, (2020), 40:63-64.
[8] Wenwen Gong. A Study on the Validity of College English Writing Test Without Outline[J]. Science and Technology Information, (2014), 12:107.
[9] Wenfeng Jia, Peixin Zhang. A Study on the Validity of the Rating Scale in Medical English Writing Test [J]. Foreign Language Testing and Teaching, (2018), 01:12-18.
[10] Li Liu. Assessing Content Quality of Writing Performance: The Hierarchical Decision-Tree Approach [J]. Modern Foreign Languages, (2013), 36:419-426.
[11] Jian Shao. On the Scoring Validity of Self-Chosen-Task Writing Test by Many-Faceted Rasch Model [J]. Journal of Xi'an International Studies University, (2021)29:72-77.
[12] Jianxiong Shi. Study on the Construction of a Scoring Scale for the Diagnostic Writing Test[J]. Applied Linguistics, (2022)03:114-123.
[13] Jin Tian. Backwash Effect of Writing Test on Promoting College English Teaching and Learning[J]. Journal of Guangxi Minzu University, (2016)38:165-168.
[14] Xuefeng Wu. Establishment and Validation of a Rating Scale Model for English Writing Tests[J]. Foreign Languages and Literature, (2018)34:137-146.
[15] Yue Wu. Validation of a Reading-to-Write Integrated Test Task Based on Multi-Facet Rasch Measurement [J]. Technology Enhanced Foreign Language Education, (2017)01:55-61.
[16] Ziwei Wang and Li Lu. Study on the Backwash Effect of College English Writing Test on Teaching [J]. Comparative Study of Cultural Innovation, (2020)4:35-37.
[17] Rui Xia. The Backwash Effect of Writing Test on College Entrance Examination English Writing Teaching [J]. Overseas English, (2018)20:109-110.
[18] Wei Xiao and Xiangdong Gu. A Comparative Study on the Backwash Mechanics of CET-6, IELTS and TOEFL Writing Modules [J]. Foreign Language Learning Theory and Practice, (2022)03:94-103
[19] Wanyi Xie. A Study on the Content Validity of Writing Test in PRETCO [J]. Journal of Changchun Finance College, (2013)04:52-54.
[20] Zongrui Xu. Negative Backwash Effects of Writing Tests of China' High-Stake English Examinations [J]. Journal of Yangzhou University, (2018)22:115-119.
[21] Zhiqiang Yang. A Study on the Grading Deviation of Adult English writing test for Bachelor Degree [J]. Education and Examinations, (2018)01:54-59.
[22] Fen Zhang. The Backwash Effect of CET-4 on College English Teaching [J]. English Square, (2017) 08: 102-103.
[23] Hong Zheng. The Negative Backwash Effect of TEM-4 Writing Test on EFL [J]. Foreign Investment in China, (2013)18:340.
[24] Haifeng Zhang. Validity Analysis of Writing Test [J]. Tea in Fujian, (2018)40:254.
[25] Peixin Zhang and Wenfen Jia. An Investigation into the Cognitive Validity of the TEM-8 Integrated Writing Test [J]. Modern Foreign Languages, (2021)45:257-269.
[26] Shaoyan Zou and Jinsong Fan. A Preliminary Examination of the Rating Scale Validity of the CET-4 Writing [J]. Foreign Languages and Literature, (2019)35:148-156.
[27] Shaoyan Zou and Jinsong Fan. Pinpointing Analytic Rating Criteria for EFL Writing Assessment from Raters' Perspectives [J]. Contemporary Foreign Language Studies, (2022)04:133-143.