Lecture Notes on Language and Literature (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/langl.2024.070804
ISSN 2523-5869 Vol. 7 Num. 8

A Scoping Review of Research (2000-2024) on Washback in China

Mengyuan Shangguan

Xi'an International Studies University, Xi'an, China

Keywords: Washback, validity, TEM-8

Abstract: The backwash effect is an important component of the post-exam impact and is one of the criteria for evaluating the quality of exams. This study analyzes papers published in 14 foreign language journals on CNKI (China National Knowledge Infrastructure) between 2000 and 2024. Using CiteSpace software for bibliometric and visual analysis, the study explores high-frequency keywords, keyword clusters, and keyword bursts. The findings reveal the following characteristics of related research: The publication timeline shows a wave-like pattern. Research on the backwash effect mainly focuses on the following hot topics: the backwash effect and validity, the backwash effect of the TEM-8 exam, and empirical studies on the backwash effect. The fields of testing, speaking tests, the TEM-8 exam, etc., are at the forefront of backwash effect research. The study also finds several issues within the backwash effect research, including repetitive research subjects, an uneven distribution of exam types, relatively simplistic research methods, and overgeneralization. The paper further suggests future research directions and provides references for scholars conducting related studies.

1. Introduction

Washback refers to the "impact of testing on teaching and learning" [1]. It is an inseparable part of the construct validity of exams [2], an important component of post-exam effects, and one of the criteria for evaluating exam quality [3]. In recent years, Chinese research on washback has deepened, with many scholars conducting extensive theoretical and empirical studies on the topic. CiteSpace, through a series of visual maps based on a large body of literature, can accurately and objectively present the overall development dynamics of the field. Therefore, this study analyzes papers published in 14 foreign language journals on CNKI (China National Knowledge Infrastructure) between 2000 and 2024, using CiteSpace software for bibliometric visualization. The analysis focuses on high-frequency keywords, keyword clusters, and keyword bursts to explore the characteristics of publication timelines, research hotspots, research themes, and research frontiers in the field of backwash effect studies, providing references and insights for scholars conducting related research.

2. Data Source

The data for this study comes from the CNKI (China National Knowledge Infrastructure) database. In CNKI, an advanced search was performed using the keywords "backwash effect" "washback" and

"post-exam effects" with the search period set to "2000 to July 2024". The search was conducted across Chinese foreign language journals. To ensure the relevance of this study, the selected papers were cross-checked against the CNKI publication database. After manual screening to remove irrelevant literature, a total of 80 valid research papers were returned and used as the data source for this study.

3. Results and Analysis

3.1. Analysis of Publication Timeline

The author analyses the time distribution of publication volume in Chinese research on washback, which can clearly reveal its historical development characteristics. This paper uses Excel to create a time distribution graph of the publication volume (see Figure 1).

From Figure 1, it can be observed that over the past two decades, Chinese research on washback has shown a wave-like development trend, which can be roughly divided into three stages. From 2000 to 2011, the research was in its initial development stage, during which the publication volume showed a slow growth trend. From 2012 to 2014, there was a significant increase in publication volume, peaking at 11 papers in 2014, indicating that scholars' attention to this field had risen. Since 2015, the research has entered a stage of steady development, during which the publication volume has remained at a relatively stable level.

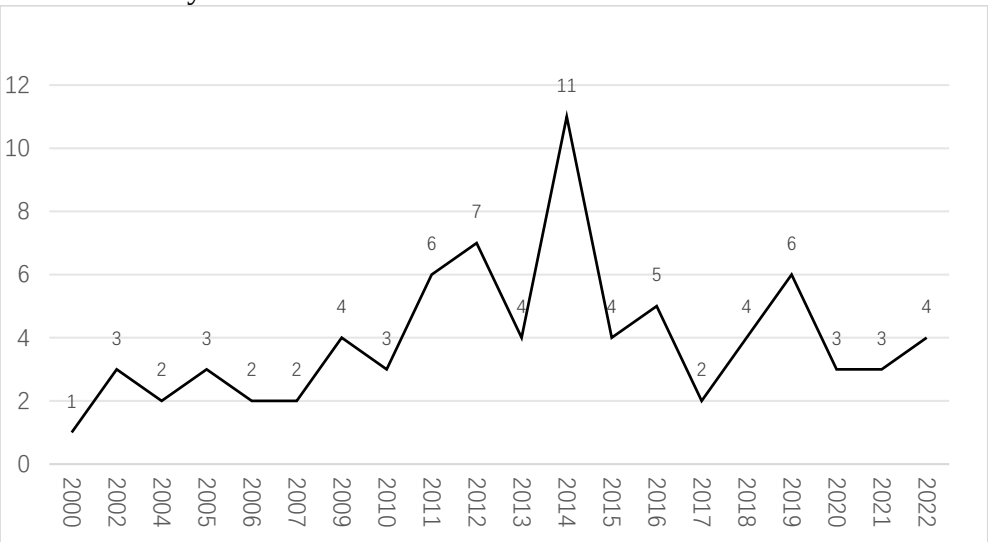

Figure 1: Statistics on the number of published papers (2000-2024)

3.2. Analysis of Research Hotspots

The core content in the literature revolves around keywords, which reflect the research hotspots. Therefore, the author uses CiteSpace software to count the frequency of keyword occurrences. The time slices are set to 1 year, with the node type selected as "Keywords" and other parameters set to default. This results in a co-occurrence knowledge map of keywords with 112 nodes, 168 links, and a density of 0.027. To further assess the importance of each keyword in the network graph, the author also calculates the high centrality keywords based on the frequency statistics. After organizing the data, the top 10 high-frequency keywords and high-centrality keywords are summarized in a table (see Table 1). The higher the frequency and centrality of a keyword, the more important the node is in the field of language testing research [4].

To further clarify the research hotspots in domestic studies on the backwash effect, the author

reviews and analyzes the 10 high-frequency keywords. Excluding the search keywords “backwash effect” “washback” and “post-exam effects”, the study identifies three main research hotspots: the backwash effect and validity research, the backwash effect of TEM-8 exam and empirical research on the backwash effect.

Table 1: High-frequency word statistics

Rank	High-Frequency Keywords		High-Centrality Keywords	
	Keywords	Frequency	Keywords	Frequency
1	Validity	7	Validity	0.38
2	TEM-8 Exam	4	TEM-8 Exam	0.19
3	Exam participants	3	College Entrance Exam	0.13
4	Empirical Research	3	Two Exams a Year	0.08
5	Listening and Speaking Test	3	Exam Reform	0.08
6	College English	3	Empirical Research	0.06
7	College Entrance Exam	2	Listening and Speaking Test	0.06
8	Fairness	2	Fairness	0.06
9	Longitudinal Research	2	English Teaching	0.04
10	High School Entrance Exam	2	High School	0.04

3.2.1. Washback and Validity Research

Research on washback cannot be separated from the exploration of its relationship with exam validity [5]. There are three main viewpoints in academia regarding the relationship between washback and validity. The first viewpoint suggests that washback of exams is closely related to validity, using the washback effect as an important criterion for evaluating exam validity. Messick argued that washback is an inseparable part of the construct validity of exams, emphasizing that test designers must consider the potential impacts that exams may produce, i.e., the washback effect. The second viewpoint holds that there is no inherent direct connection between washback and validity. Alderson and Wall viewed the washback effect as a complex phenomenon and argued that it should not be directly linked to exam validity. Qi Luxia [6] expressed agreement with the view that the washback effect is the most important indicator for evaluating language exams but disagreed with including it as part of exam validity. The third viewpoint acknowledges the differences between the washback effect and validity, while also affirming their close connection. Chen Xiaokou [7] argued that although the definitions, research subjects, and scope of the two concepts are not the same, the washback effect and validity are closely related. This is because exam factors, which are the main contributors to the washback effect, also involve key elements of validity research, such as exam methods, content, test design, and scoring.

The author believes that washback, as an indicator of validity, can provide clues about the validity of a test, especially in terms of construct validity. For example, if a test generates significant washback effects on teaching and learning, and these effects align with the intended goals of the test (such as improving language skills, particularly in the areas the test is designed to assess), this suggests that

the test may be valid. On the other hand, if the washback effects lead to teaching practices that are inconsistent with the intended goals of the test (such as overemphasizing certain content while neglecting the comprehensive scope of the test), this may indicate validity issues with the test. For example, if a test designed to assess reading ability causes teachers to overemphasize vocabulary instruction while neglecting the development of critical reading strategies, this suggests that the test may not effectively measure the full scope of reading proficiency, thereby impacting its validity.

3.2.2. Washback of TEM-8 Exam

Many Chinese scholars have focused on the washback effect of TEM-8 exam. For example, Zou Shen [8] examined the impact of the TEM4 and TEM8 exams on university English teaching and learning from three perspectives: exam administrators, teachers, and students. They concluded that to ensure a positive washback effect, research on the washback effect should be integrated throughout the entire curriculum design process. In this interactive and integrated process, TEM- 8 exam is more likely to have a positive impact on teaching and learning. Xu Qian [9] conducted a survey of several foreign language experts and English department heads regarding the washback effect of TEM-8 exam. The results indicated that respondents generally held a positive view of TEM-8 exam, believing that its positive impact on English major teaching outweighed its negative effects.

The author believes that the washback effect of the TEM-8 exam can be both positive and negative. Positive washback includes a focus on proficiency, as students improve skills in listening, speaking, writing, and translation. Universities align their curricula to better prepare students, fostering a more focused and rigorous language education. The challenging nature of the exam also motivates students to enhance their abilities, raising overall proficiency. Negative washback includes an overemphasis on test preparation, leading to rote learning and neglecting a well-rounded education. The focus on reading, writing, and translation may limit attention to oral communication and cultural understanding. High stakes can also cause stress and anxiety.

3.2.3 Empirical Research on Washback

Empirical research on the washback effect in domestic language testing began in the late 1980s. Li Xiaoju [10] was the first to conduct empirical research on the washback effect in China. He pioneered a study on the washback effect in high school English teaching and conducted preliminary empirical research on its impact in China. The survey results found that the English test for the college entrance examination had achieved some expected effects on both high school English teaching and learning. Gu Xiangdong [11] employed research methods such as classroom observation, discourse analysis, and interviews to conduct a longitudinal study on the washback effect of the reform of the CET-4 and CET-6. The study found that, after the CET reform, the thematic patterns of college English classroom teaching had remained largely unchanged, but significant changes had occurred in teaching plans, content, methods, and other area. Overall, Chinese empirical research has mainly focused on the impact of the CET on teaching and learning. Regarding the nature of the washback effect, the research found that the CET has both positive and negative washback effects.

3.3. Analysis of Research Topics and Frontiers

3.3.1. Analysis of Research Topics

To focus on the research topics of the washback effect from 2000 to 2024, the author performed a keyword clustering operation to generate a cluster label map. Different thematic clusters can represent the main research areas of each group, thereby clearly presenting the core topics of the research [12]. In the keyword clustering results for research on the washback effect, seven cluster labels were

generated. These labels are as follows: #0 CET, #1 Speaking and Listening Tests, #2 Validity, #3 Language Testing, #4 TEM-8 Exam, #5 High School, #6 Review and Prospect.

To roughly outline the core academic landscape of the research and minimize redundancy within the keyword nodes of each cluster (see table 2), the author conducted a statistical analysis of the node information for each cluster. The key information for each cluster was then further refined by removing redundant and repetitive content, and the node information was summarized and generalized for greater precision and accuracy. This process aimed to provide a solid foundation for future research. The relevant research topics identified include: the washback effect of CET, the washback effect of speaking and listening tests, the relationship between washback and validity, and the washback effect of the Gaokao English exam.

Table 2: Cluster label statistics

Serial No.	Cluster Label	Main Content
#0	CET	Washback effect, longitudinal studies, College English Exam, CET-4 online exam, college English teaching, listening tests, teaching
#1	Listening and Speaking Test	College Entrance Exam English, washback effect, second language acquisition, TEM-8 writing, grading criteria, Shanghai English entrance exam
#2	Validity	English education, score interpretation and usage, test design, data feedback
#3	Language Testing	Validity verification, effective teaching, assessment models, sociology of language testing
#4	TEM-8 Exam	English major level 4, teaching quality assessment, humanities courses, authority
#5	High School	High school English, English teaching, listening exams, washback effect
#6	Review and Outlook	Classroom teaching, college English, development trends, empirical research

3.3.2. Frontier Analysis of Research

To grasp the development trends and future research directions of washback effects, and to understand the frontier of this field, the author applied the burst word detection algorithm in CiteSpace software to analyse keywords and detect burst words along with their distribution map. A total of 12 burst words were detected, and they are: Post-effect, Test, Longitudinal Study, Exam Stakeholders, Language Testing, Empirical Research, English Teaching, High School, Validity, Listening and Speaking Test, Washback Effect, TEM-8.

The analysis revealed that the time spans of the burst words varied. The word with the longest duration of burst was “test”, which started in 2005 and showed a decline starting in 2011. The top three burst words were “post-effect” “test” and “listening and speaking test”, which became burst words in 2005 and 2018, respectively. In recent years, there has been considerable research on the washback effect of listening and speaking tests. From the map, it is evident that research on listening and speaking tests as well as the TEM-8 exam began in 2019 and 2020, respectively, and has continued to the present. This indicates that, both currently and in the foreseeable future, research on “listening and speaking tests” and the “TEM-8 exam” will remain a key frontier and focus for scholars.

4. Conclusion

The related research on the washback effect presents the following characteristics: Firstly, the publication timeline shows a wave-like pattern, with research divided into three main stages. Secondly, research on the washback effect mainly involves the following hotspots: the relationship between the washback effect and validity, the washback effect of the TEM-8 exam, and empirical research on the washback effect. Thirdly, through keyword clustering, the research topics in the field of the washback effect focus primarily on the washback effect of the CET, the washback effect of speaking and listening tests, the relationship between washback and validity, and the washback effect of the Gaokao English exam.

From these conclusions, it is evident that Chinese research on the washback effect is gradually maturing and has yielded rich results. It can be inferred that the number of papers will continue to increase steadily in the coming years. However, there is still significant potential for improvement in the field of washback research, which can be summarized in the following four points: (1) Repetition in research subjects. (2) Uneven distribution of research exam types. (3) A relatively narrow range of research methods. (4) Simplification and overgeneralization of washback research. In response to these shortcomings, the author offers some prospects for the development of research in the field of washback:

First, expand the range of research subjects. In addition to teachers and students, research could be conducted on other stakeholders such as parents, textbook writers, and education administrators. Second, study multiple types of exams. For example, research could be conducted on the washback effect of school-based exams, the CATTI (China Accreditation Test for Translators and Interpreters), graduate entrance exams, the middle school English exam, the National Public English Test, and the College English Application Test. Third, diversify research methods. Single research methods may not be sufficient to prove the reliability of results. A combination of different tools (such as surveys, in-depth interviews, document analysis, think-aloud protocols, etc.) should be used to complement one another, verify data from multiple perspectives, and conduct triangulation to ensure the reliability and validity of the research.

Study the washback effect from multiple perspectives. Research could be conducted from the perspective of cognitive psychology, focusing on the role of stakeholders' awareness and beliefs in the formation of washback effects. This approach would provide a more scientific and rational explanation of the washback phenomenon and its mechanisms. Additionally, interdisciplinary research, such as the use of think-aloud protocols or eye-tracking experiments, could offer new insights.

In conclusion, in conclusion, future research on washback effects in China should focus on a broader and more multidimensional exploration. In addition to traditional studies involving teachers and students, other educational stakeholders, such as parents, textbook writers, and education administrators, should be included to gain a comprehensive understanding of how different perspectives influence washback. Moreover, as the types of exams diversify, future studies should expand to include various exams, such as professional certification tests and vocational assessments, to evaluate the broader impact of exams on education. The integration of multiple research methods will enhance both the depth and breadth of washback research, promoting further development in both theoretical and practical aspects of the field.

References

[1] Alderson, J.C. & Wall, D. (1993). Does washback exist? Applied Linguistics, 14(2), 115-129.
[2] Messick, S. (1989). Validity. In R.L. Linn (Ed.), Educational Measurement, 3rd edition (pp. 14-22).
[3] Qi, Luxia (2012). Recent research and future prospects of the washback effect of language testing. Modern Foreign

Languages, 35(2), 202-208+220.
[4] Guo, Jia (2020). An investigation of the authenticity of IELTS speaking based on communicative language testing theory. China Examinations, (6), 19-26.
[5] Dong, Manxia (2019). Several basic issues that need clarification in the study of washback effects in language testing. Foreign Language Teaching Theory and Practice, (03), 50-57.
[6] Qi, Luxia (2011). Theoretical and empirical research on the washback effect of language testing. Foreign Language Teaching Theory and Practice, (04).
[7] Chen, Xiaokou (2016). The reproductive effect of language testing research on validity studies. Foreign Language Testing and Teaching, (01), 39-46.
[8] Zou, Shen & Dong, Manxia (2014). Twenty years of washback effect research in China: Current status and reflections. Chinese Foreign Languages, 11(4), 4-14.
[9] Xu, Qian (2012). A study on the washback effect of the TEM-8 exam: A survey of foreign language experts and English discipline leaders. Foreign Language World, (03), 21-31.
[10] Li, Xiaojv, Gui, Shichun & Li, Wei (1990). The design of NMET test items and its influence on secondary school English teaching. Primary and Secondary School English Teaching and Research, (1), 1-27.
[11] Gu, Xiangdong & Peng, Yingying (2010). A longitudinal study on college English teachers' understanding of CET and its washback effect. Foreign Languages and Foreign Language Teaching, (06), 37-41+56.
[12] Zhao, Ping & Ji, Xiaoli (2016). Research on the calibration of VST test tools based on classical test theory and item response theory. Foreign Language Testing and Teaching, (2), 39-46+59.