BILINGUAL PUBLISHING GROUP

Forum for Linguistic Studies

https://journals.bilpubgroup.com/index.php/fls

REVIEW

A Systematic Review of Washback Research in China (2005–2025)

Lingxiao Li [1] , ***Tianhao Li*** [2*]

[1] *School of Foreign Languages, Qilu University of Technology, Jinan 250353, China*

[2] *Department of Foreign Languages and Literatures, Tsinghua University, Beijing 100084, China*

ABSTRACT

Washback refers to the impact of language tests on language teaching and learning. Though numerous empirical studies have been done on this topic, there is a lack of systematic reviews of these studies. In light of this, this study presents a review of washback research in the Chinese mainland over the past two decades. Specifically, 66 studies were included and analyzed in terms of the following aspects: test type, participant, research methodology, and research foci. The results showed that while the majority of the reviewed studies examined the washback of English tests, especially large-scale high-stakes ones, including the College English Test (CET), the National Matriculation English Test (NMET), and the Test for English Majors (TEM). Meanwhile, a few studies investigated the washback of non-English language tests, such as the Chinese Proficiency Test (HSK). The reviewed studies primarily recruited university students and teachers as participants, while only a few studies included other stakeholders, like school leaders and test constructors. Quantitative and mixed-methods approaches, along with cross-sectional designs, were predominantly adopted, while qualitative research methods and longitudinal research designs were comparatively less used. Moreover, three major research themes were identified: washback effects on language learning, washback effects on language teaching, and the factors influencing these effects. This review maps the current state of washback research in the Chinese mainland and offers directions for future research.

Keywords: Washback; Language Testing and Assessment; Systematic Review

*CORRESPONDING AUTHOR:

Tianhao Li, Department of Foreign Languages and Literatures, Tsinghua University, Beijing 100084, China; Email: tianhaoli21@163.com

ARTICLE INFO

Received: 29 July 2025 | Revised: 13 August 2025 | Accepted: 9 September 2025 | Published Online: 22 September 2025
DOI: https://doi.org/10.30564/fls.v7i10.11215

CITATION

Li, L., Li, T., 2025. A Systematic Review of Washback Research in China (2005–2025). Forum for Linguistic Studies. 7(10): 71–84.
DOI: https://doi.org/10.30564/fls.v7i10.11215

1. Introduction

Language tests can have a profound influence on language teaching and learning[1–3]. Several terms were used to describe such effects of testing, like *'washback'*[1], *'backwash'*[2], *'test impact'*[3], and *'consequential validity'*[4]. Among them, washback is the most accepted and well-known term[5]. In the mid-1990s, Alderson and Wall pointed out the lack of empirical research on this widely perceived phenomenon[1] and proposed a question: 'Does washback exist?'" (p. ix). Since then, researchers worldwide have conducted numerous empirical studies on the washback effects of language tests[6–8]. They have made great efforts to verify its existence and attempted to answer questions, such as what kind of washback effects a test may produce and what factors may influence such effects[6–8].

So far, washback has remained a key topic in language testing research[5,6]. Meanwhile, the Chinese mainland has a large-scale and unique education system, and the diverse language tests within it, such as NMET and CET, provide favorable research conditions for exploring the washback of language tests on language teaching and learning[9–14]. According to a bibliometric analysis of 243 washback studies published from 1993 to 2023, the reviewed studies were conducted in 39 countries or regions[6]. Moreover, the Chinese mainland has demonstrated consistent and remarkable growth in the number of washback studies, surpassing all other countries and regions[6]. Despite the substantial volume of washback studies carried out in the Chinese mainland, there were few systematic reviews of these studies. The disparity between the large-scale empirical exploration and the scarcity of systematic reviews calls for a comprehensive synthesis of existing research and directions for future washback research. Given this, this study provides a systematic review of washback research in the Chinese mainland over the past two decades (from 2005 to 2025). By investigating the unique trends and patterns of washback research in the Chinese mainland, this review will not only fill the existing gap in the literature but also contribute to a more comprehensive understanding of the washback effect in the Chinese context.

2. Literature Review

2.1. Theoretical Models of Washback

Researchers have proposed various theoretical models and frameworks of washback[15–17]. Alderson and Wall's Washback Hypotheses laid the groundwork for later empirical studies on this topic[5,18]. In 1993, they proposed a series of hypotheses about the washback effects of language tests[1]. They believed that a test can influence language teaching (e.g., what and how teachers teach) and learning (e.g., what and how learners learn). Also, whether the test has washback depends on its importance, and tests have different effects on different teachers and learners[1]. Inspired by their re-conceptualization of washback, researchers began to critically think about the notion of washback[15–17]. However, these hypotheses reflect a rather general and simplistic understanding of washback and do not mention that washback is a social phenomenon[19].

Hughes proposed a trichotomy model of washback, which includes "participants-process-product"[15]. Hughes emphasized the necessity of distinguishing these three parts and suggested that they were all influenced by the test. A test would have an impact on the participants' (e.g., teachers, students) perceptions and attitudes, which would, in turn, affect the language teaching and learning process and ultimately affect the product (e.g., language learning outcomes). Hughes listed the necessary conditions for the occurrence of the washback, such as the importance of the test to participants, teachers' desire for students' success in learning, and the available resources for preparing the test. Hughes's trichotomy model clearly reflects the mechanisms of washback. However, it lacks the consideration that for some participants, the process might not directly lead to significant positive products, such as improved language skills, but might produce some ancillary outcomes. Besides, this model only shows a one-way relationship between the three aspects. It overlooks the fact that testing is not a purely academic activity, but rather takes place in social contexts[19].

Based on Alderson and Wall's work and Hughes' model, Bailey introduced another model of washback. This model

emphasizes the mutual interaction between a test and the potential participants [16]. In the model, a test can affect four categories of participants (i.e., students, teachers, designers of textbooks and curricula, and academic scholars) and produce corresponding outcomes. Bailey put forward the term "washforward", suggesting that these participants can also impose influences on the test [16]. Bailey also distinguished between the ideas 'washback to the learners' (the direct effects on the test-takers) and 'washback to the programme' (the impact on other relevant stakeholders, like teachers, school leaders, and curriculum designers) [16]. These two terms were both consistent with the hypotheses of Alderson and Wall regarding the impact of tests on teaching and learning [1]. Apart from the models mentioned above, there are many well-developed theoretical models [17,20–23]. Guided by these theoretical models and relevant discussions on washback, researchers have carried out numerous empirical studies [6].

2.2. Empirical Research on Washback

Early empirical research has predominantly centered on the washback of two types of language tests: global high-stakes tests (e.g., the International English Language Testing System (IELTS) and Test of English as a Foreign Language (TOEFL)), and national or regional campus-based tests or entrance tests (e.g., the NMET and CET in China) [23–26]. In terms of studies on global language tests, Alderson & Hamp-Lyons focused on English teachers and English as a Foreign Language (EFL) learners in the USA and found that the TOEFL influenced the teachers' teaching content and methods [24]. Green recruited Chinese university English teachers and students to explore the washback of IELTS on them [27]. As for research on national or regional tests, Qi explored the washback effect of the NMET on English teaching and learning in Chinese high schools. The study involved students, their English teachers, and the test constructors of the NMET [10].

Despite the predominant attention on English tests [6], there has been a growth in research on the washback of non-English language tests, such as the Chinese language proficiency test (HSK) [28–30] and the Portuguese language proficiency test (CAPLE) [31]. For instance, Kong and Zhang conducted a questionnaire survey on 1616 Chinese as a Second Language (CSL) learners from 41 countries and interviewed 6 of them [29]. They found that the HSK significantly influenced the learners' Chinese learning processes and outcomes. Liu concentrated on 102 Chinese undergraduates who majored in Portuguese as well as 28 Portuguese teachers, and found positive washback of the test on the participants, such as increased language abilities and learning confidence [31].

A significant body of research has confirmed that tests exert notable washback on both language learning (e.g., influencing learning outcomes, strategies, and attitudes) and teaching (e.g., promoting changes in teaching methods, content, and syllabus) [24,30–32]. For instance, Li found that NMET led to a shift in high school English teaching content from emphasizing language knowledge (grammar, vocabulary, and phonetics) to cultivating students' language abilities (listening, reading, writing, and speaking) [33]. Teachers have become more flexible in their choice of instructional materials, with an increased use of imported and self-compiled ones. Students' enthusiasm for English learning improved, and their after-school learning methods diversified. Besides, Gu found both positive and negative influences of the CET on university students and English teachers in China [25]. The test effectively promoted the refinement of the college English teaching syllabus, drew the attention of school leaders to English courses, motivated teachers and students, and enhanced students' English reading skills. Meanwhile, it also led to the "teaching to the test" phenomenon due to an overemphasis on passing rates. He summarized that the positive effects outweigh the negative ones.

Scholars have identified various factors that can influence the washback effects of language tests [22,34]. For instance, in Alderson & Hamp-Lyons' study, the teacher participants in the TOEFL classes and in the normal English classes experienced different degrees of washback [24]. In Qi's study on the washback of the NMET, he compared the English teaching practices in Chinese secondary schools at that time and the intentions of test designers [10]. He pointed out that due to the conflict between school-based English teaching and the intentions of test constructors, the test had limited intended washback. While the English classrooms still emphasized learning language knowledge, the test designers wanted to use the test to promote the communicative use of English. Shohamy et al. found that high- and low-stakes language tests (Arabic and English) exerted different washback effects on university students in Israel. The high-stakes test had more intense washback than the low-stakes test [34]. Shih studied English major university students in

Taiwan and found that the General English Proficiency Test (GEPT) had limited effects on their English learning [22]. He attributed the insignificant washback in the study to the participants' majors. Based on the empirical findings, he proposed a washback model that includes three major categories of factors jointly affecting the washback effects on students' learning and psychology. The three types of factors were intrinsic (e.g., individual differences), extrinsic (e.g., socioeconomic factors), as well as test-related factors (e.g., stakes and importance of the test).

Given the large number of empirical studies on washback, some scholars have carried out systematic reviews of these empirical investigations [8,35–37]. For example. Allen and Tahara conducted a systematic review of washback research in Japan [36]. They reviewed 32 empirical studies published before 2021 and examined them in terms of several aspects, including test type, participant group, research methodology, and so on. Chen focused on washback research in China and analyzed 38 washback studies published in Chinese journals, examining the test type, research methodology, and research focus of each study [8]. These reviews help reveal the trends and patterns of washback research. Besides, the number of washback studies conducted in the Chinese mainland has been increasing steadily since the 2000s [6,8]. Yet, there are relatively few systematic reviews of these studies. To fill this gap, this study focuses on the empirical studies conducted in the Chinese mainland over the past two decades and provides a systematic review. This study can contribute to an in-depth understanding of the washback phenomena in the Chinese context. It seeks to answer the following research questions:

(1) What are the characteristics of washback studies (test type, participant, research methodology) conducted in the Chinese mainland from 2005 to 2025?
(2) What are the research foci of these studies?

3. Method

3.1. Literature Search

The literature search and screening followed the guidelines of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) (See **Figure 1**) [38]. Articles were searched in three databases: the Web of Science (WOS), Scopus, and CNKI (China National Knowledge Infrastructure). For the two English-language databases, WOS and Scopus, the search keywords were "washback" AND "China". As for the Chinese-language database, CNKI, we used "反拔" (washback in Chinese) as the search keyword. After obtaining the search results from these databases, we filtered them to include only articles published between 2005 and 2025. After the initial search, duplicate articles were removed. The literature search was conducted independently by the first author. A total of 969 articles remained for further screening.

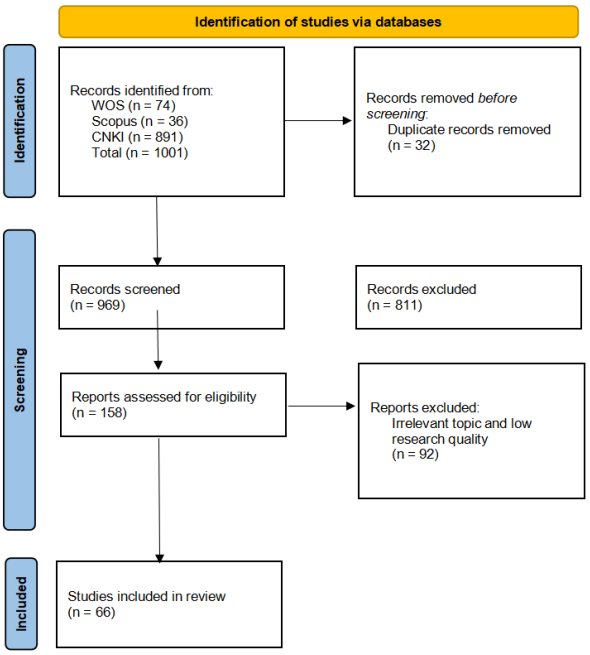

Figure 1. Selection procedure.

3.2. Literature Screening

During the screening process, specific inclusion and exclusion criteria were applied. The inclusion criteria were as follows: The studies should be empirical research that focused on the washback of language tests and were conducted in the Chinese mainland. The articles needed to be published in either English or Chinese, and their full-text versions should be accessible online. The studies must involve language teachers or learners in China, and focus on language tests used in Chinese educational settings. Only peer-reviewed journal articles were included. The exclusion criteria were as follows: First, works such as literature reviews, theoretical papers, and book reviews were not consid-

ered for this review. Second, studies published in languages apart from English or Chinese were excluded. Additionally, any research conducted outside the Chinese mainland was excluded. Furthermore, after the initial screening, the quality of the included studies was evaluated. Specifically, each study was carefully examined in terms of the research questions and research methodology. Studies with irrelevant research topics or low research quality were removed. For example, in quantitative research, studies with a small sample size or unreliable measurement tools were excluded. The two authors first independently conducted the screening of the literature and subsequently compared their respective results. Discrepancies were resolved through discussion to reach a consensus. Finally, 66 articles were left for further in-depth analyses.

3.3. Data Analysis

Firstly, relevant information was extracted from the selected articles, including the publication year, author, test type, participants, research methodology, research questions, and major findings. To answer the first research question, quantitative analyses were carried out on the data. All the results were presented in figures and tables for clear visualization. Specifically, the researchers counted the test types that were focused on in each study and calculated the frequencies of various test types. In addition, regarding the participants, they identified different participant types in each study and calculated their frequencies. Moreover, in terms of the research methodology, the researchers counted the frequencies of different research designs (whether the research was cross-sectional or longitudinal), research instruments, and research methods (quantitative, qualitative, or mixed-methods) employed in the reviewed studies.

To answer the second research question, a thematic analysis was conducted, which involved two rounds of coding. In the first round, broad and initial themes were generated and matched with corresponding data excerpts that could serve as evidence to illustrate them. Through this process, three main themes were identified: "washback effects on language learning", "washback effects on language teaching", and "factors that influence these effects". In the second round of coding, the sub-themes were identified based on the broad themes generated in the first round. For example, under the theme of "washback effects on learning", sub-themes such as "effects on the learning process" and "effects on the learning outcome" were further determined. To ensure the reliability of the coding results, the two researchers carried out the coding work independently and then compared the coding results. The discrepancies in the coding results were addressed through a series of discussions.

4. Results

4.1. Test Type

As shown in **Figure 2**, 84% of the reviewed studies explored the washback effects of English tests (N = 56). Among these tests, CET, NMET, and TEM were the most frequently studied tests. Specifically, 21 studies focused on the washback of CET, 14 studies on that of NMET, and 10 studies on that of TEM. In addition, several studies examined the washback of international standardized English tests, including TOEFL (N = 3), IELTS (N = 3), and PTE (N = 1). Moreover, six studies examined the washback of school-based English tests, such as those designed by specific universities.

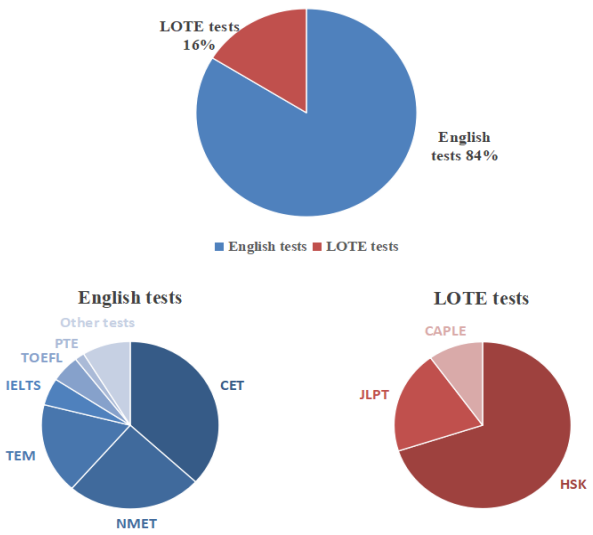

Figure 2. Distribution of test type.

In contrast, 16% of the reviewed studies (N = 11) explored the washback of non-English language tests. Specifically, eight studies examined the effects of the HSK, two focused on the Japanese Language Proficiency Test (JLPT), and one centered on the CAPLE. Notably, only one study made a comparison of the washback effects of different tests [9]. In the remaining studies, researchers all focused on the washback of a particular test. The results indicate that researchers

have paid significantly more attention to English tests, particularly the three large-scale domestic English tests —CET, NMET, and TEM —than to other types of tests.

4.2. Research Participants

Figure 3 and **Table 1** illustrate the distribution of participant types in the reviewed studies. As presented in **Figure 3**, 51 studies (82%) focused on Chinese university students and teachers, while 11 studies recruited participants from Chinese high schools (18%).

As shown in **Table 1**, among the 51 studies carried out in Chinese university settings, 29 studies recruited only university students as participants, while 14 studies involved both university teachers and students. Furthermore, four studies focused solely on university teachers, four studies recruited teachers, students, and other stakeholders, such as test constructors. Two studies concentrated solely on the relevant stakeholders other than university teachers and students, and two studies recruited both university teachers and other relevant stakeholders. As for the 11 studies that involved participants from Chinese high schools, 6 studies only recruited high school students as participants, while 3 studies solely concentrated on high school teachers. Moreover, one study recruited both high school teachers and students, while another study involved teachers, students, and other relevant stakeholders. These results suggest that so far, researchers have paid significantly more attention to Chinese university teachers and students compared to high school teachers and students. Also, students emerged as the most frequently studied participant type. Furthermore, apart from teachers and students, other relevant stakeholders, like school leaders and test designers, have received relatively little research attention.

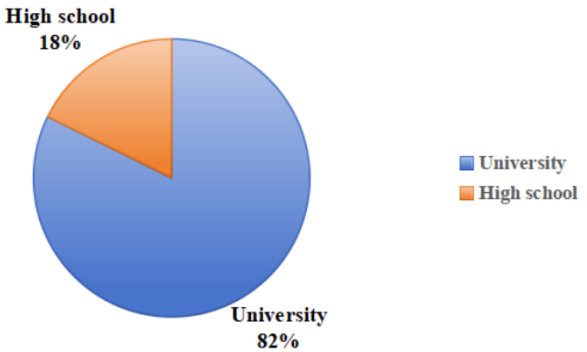

Figure 3. Distribution of participants.

Table 1. Distribution of participants.

Participant Type	N
University	55
Student	30
Teachers and students	13
Teacher	4
Teachers, students, and other stakeholders	4
Other stakeholders	2
Teachers and other stakeholders	2
High school	11
Student	6
Teacher	3
Teachers and students	1
Teachers, students, and other stakeholders	1

4.3. Research Methodology

In terms of research methods, as illustrated in **Figure 4**, 36 studies (55%) adopted mixed-method approaches, 27 studies (41%) used quantitative methods, and only 3 studies (4%) used qualitative methods. Regarding research design, 58 studies (89%) were cross-sectional, while only 8 (11%) were longitudinal. As presented in **Table 2**, among the 66 studies, 27 (41%) employed questionnaire surveys for data collection. 26 studies (39%) combined questionnaire surveys and interviews, and 10 (15%) integrated questionnaire surveys, interviews, and classroom observation. 2 studies (3%) utilized interviews and journals, and 1 (2%) only used classroom observation to collect data. These results suggest that researchers tend to employ quantitative and mixed-methods approaches, as well as cross-sectional designs, more frequently than qualitative methods and longitudinal designs. Questionnaires were the most frequently utilized instrument, followed by interviews, while classroom observations and journals were less frequently used.

4.4. Research Foci

In terms of the research foci of the reviewed studies, three major themes and their sub-themes were identified. The first theme, washback effects on language learning, was categorized into three sub-themes: effects on language learning processes, effects on language learning outcomes, and effects on students' psychology. Regarding the effects on learning processes, according to the reviewed studies, the language test significantly influenced students' learning behaviors (e.g., test preparation, learning activities)[29,30]. As for

the impacts on language learning outcomes, some researchers found that the test had an impact on language learning outcomes, such as students' test scores and self-rated language achievement[11,12,39]. As for the effects on students' psychology, the reviewed studies demonstrated that the language test influenced students' attitudes towards the test, language learning motivation, strategy, anxiety, etc.[32,40].

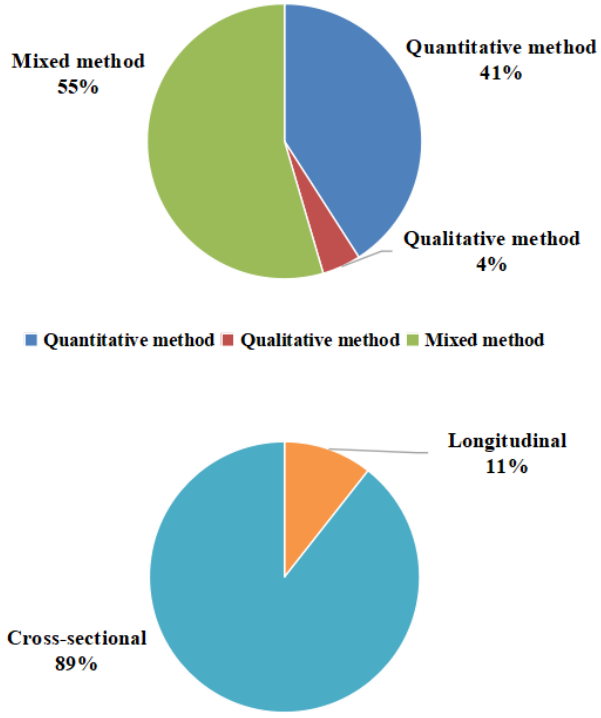

Figure 4. Distribution of research methodology.

Table 2. Distribution of the research instrument.

Instrument	N
Questionnaire survey	27
Survey + interview	26
Survey + interview + classroom observation	10
Interviews + journal	2
Classroom observation	1

The second theme, washback effects on language teaching, involves two sub-themes: effects on language teaching processes and effects on teachers' psychology. For the effects on teaching processes, the reviewed studies showed that language teachers' teaching practices (e.g., teaching content, teaching method, teaching plan) were influenced by language tests[41–44]. Regarding the effects on teachers' psychology, the studies indicated that language tests influenced teachers' attitudes towards teaching, teaching pressure, and so on[14,41].

The third major theme, factors influencing washback effects, included three sub-themes: personal, test-related, and social context factors. Personal factors involve two types of factors: student-related factors, such as language proficiency, perceptions of the test, and individual characteristics (e.g., age, gender, and grade)[41,45], as well as teacher-related factors, like years of teaching experience, teachers' perceptions of the test[46,47]. Test-related factors consist of test design, test difficulty, test stakes, and so on[48,49]. As for the social context factors, factors such as the education system, institutional policies, and management regulations can affect the washback effects of the test[41,48] (**Table 3**).

Table 3. Research foci of the reviewed studies.

Effects on language learning		
Effects on the language learning process	*Effects on language learning outcomes*	*Effects on students' psychology*
Learning practices, allocation of learning time, learning activities, etc.	Self-rated learning achievement, test scores, etc.	Attitudes towards English, learning motivation, strategy, anxiety, etc.
Effects on language teaching		
Effects on the language teaching process	*Effects on teachers' psychology*	
Teaching method, teaching goal, teaching plan, etc.		Attitudes towards teaching, teaching confidence, etc.
Factors that influence the effects		
Personal factors	*Test-related factors*	*Social context factors*
Students' age, gender, grade, language proficiency, perceptions of the test, etc. Teachers' years of teaching, perceptions of the test, etc.	Test design, test difficulty, test format, etc.	Region, education system, school regulation, etc.

5. Discussion

5.1. Characteristics of the Studies

The results showed that over 84% of the reviewed studies were concentrated on the washback of English tests. The three domestic large-scale high-stakes English tests, namely, the CET, NMET, and TEM, were the most frequently studied tests. This finding is supported by a previous review of washback research. For example, in Chen's review of washback studies in China, he focused solely on articles published in Chinese journals before 2020. The results showed that more than 70% of the studies focused on the English tests held domestically. 44.7% of the reviewed studies examined the washback of the CET, 21.2% investigated the washback of the TEM, and 13.2% concentrated on the washback of the NMET. These findings suggest that three language tests play an important role in the Chinese education system [13,39]. They can exert a profound influence on the academic and career development of Chinese students [46,49,50]. The CET is designed to assess the comprehensive English abilities of Chinese university students [49,51]. In some universities, passing this test is closely linked to students getting their graduation certificates [9]. As for the NMET, it takes up a large proportion of the Gaokao, namely the Chinese college entrance examination, and is crucial for students to enter higher education [41,46]. Cheng and Hamid noted that the NMET is far from being just an ordinary test. Instead, it is one of the most crucial competitions for Chinese students. Moreover, it can exert a long-lasting influence on Chinese education [13]. The TEM aims to test the English abilities of Chinese English major undergraduates [14]. Currently, over a quarter of a million English major undergraduates need to take this test per year [51]. Meanwhile, limited attention has been paid to international English proficiency tests, school-based English tests, as well as non-English language tests [28,29,32,48]. One possible reason that the impact of these types of tests is less significant compared to the three tests mentioned above is that these tests are not mandatory for Chinese language learners. Also, it might be difficult for researchers to seek out suitable participants.

In addition, 16% of the reviewed studies examined the washback of non-English language tests, including HSK [28,29,47], JLPT [52], and CAPLE [31]. Among these studies, the number of studies on the HSK, a Chinese proficiency test, is the largest. The test scores are a key requirement for international students applying to Chinese universities. The researchers mainly recruited international students studying in China, while some scholars focused on students at Confucius Institutes overseas. Besides, the JLPT and CAPLE are two tests that are targeted at undergraduate students majoring in foreign languages in China. For these students, the results of these two tests are of great significance to their future academic and career opportunities. Overall, these findings suggest that the scope of washback research conducted in the Chinese mainland has continued to expand. While the English test has remained the primary focus of researchers in washback research [23,25], there has been a growing interest in the washback of non-English language tests. These findings were consistent with the research conducted by international scholars [53–55]. In a review of washback studies in Japan, all the reviewed studies investigated the washback of English tests, especially those used for academic admissions, like the university entrance exam [36]. Besides, as shown in a bibliometric study of washback research, scholars have increasingly turned their attention to the washback effects of non-English language tests [31]. Yang and Jirawit investigated the impact of the HSK on Thai undergraduates [54]. They suggested that the HSK is closely linked to the personal development of Thai university students, helping them meet their academic and professional needs and providing more opportunities in both areas [55]. In Laotrakunchai and his colleagues' study on high school students and Chinese teachers in Thailand [55], the results showed that the test significantly influenced the student participants' Chinese learning. It has enhanced their learning interest, boosted their learning confidence, and improved their Chinese language abilities.

Regarding participants of the reviewed studies, most studies recruited students and teachers from Chinese universities. In contrast, a few studies included participants from Chinese high schools. This finding is consistent with the reviews of washback research conducted by Chen [6] and Xie and Jia [8]. Xie and Jia suggested that most washback research is centered on language tests at the tertiary education level and the "gate-keeping" tests during the transition from secondary to tertiary education. This finding aligns with the findings of previous reviews of washback research. In Allen and Tahara's review of washback studies in Japan, more than half of the reviewed studies recruited participants from

Japanese universities, while others focused on participants from high schools [36]. The concentration on university and high school populations in washback research is understandable. Students and teachers are the two key stakeholders in language teaching and learning [1–3], and it is more convenient for researchers to collect data from them. However, there is a notable lack of research on the washback effects of language tests on students and teachers of other educational levels, like primary, middle school, and vocational school students and teachers. These language learners and their teachers are in different educational settings from those in universities and high schools, and the language teaching and learning situations can be quite different. The washback effects of language tests on these participants may differ from current research findings, and further investigation is necessary.

Furthermore, though students and teachers were the most commonly studied types of participants, some scholars have recruited other stakeholders, such as school leaders and test constructors, as a supplement [10,51]. For example, Fan and his colleagues focused on the effect of the TEM on teachers' perceptions [51]. They recruited 194 English teachers for a questionnaire survey and conducted interviews with 16 of them. Additionally, they interviewed three members of the TEM committee. These three members had worked on the committee for more than a decade and were responsible for making decisions on the design and reform of the test. In Zhang and Kong's study of the washback effects of the HSK, 1616 students, 112 teachers, and 41 stakeholders were involved [30]. The student participants were Chinese as a second language (CSL) learners. The teacher participants were engaged in teaching Chinese as a second language at universities. The researchers also interviewed 28 school leaders from Chinese universities and 13 administrators of overseas Confucius Institutes. This finding is supported by Chen's review of 38 empirical studies conducted in the Chinese mainland [8]. The researcher found that 34 of these studies examined the washback of the test on students and teachers. In contrast, only four studies involved other stakeholders, such as school leaders and test designers, as participants. Many scholars have noted that washback has the potential to impact not only individuals but also the educational system [5,7,13]. Although the number of such studies is relatively limited, they have significantly deepened our understanding of the washback effects of language tests by integrating the perspectives of various stakeholders.

Regarding the research methodology, this study found that the reviewed studies predominantly employed quantitative and mixed-methods approaches. Quantitative research is usually based on the analysis of large-sample data, and can provide clear numerical evidence of the topic under investigation [56]. Combining both quantitative and qualitative data, mixed-method approaches offer a comprehensive perspective on the research questions and facilitate the triangulation of findings [57]. Qualitative approaches focus on exploring in-depth and rich data related to the topic. However, among the reviewed studies, only two used qualitative methods exclusively. Questionnaire survey was the most frequently used tool for data collection, followed by interviews, while classroom observations and journals were less commonly utilized. This may be because questionnaires are cost-effective and can be conducted more rapidly [58].

Additionally, most of the reviewed studies employed cross-sectional designs, while a few adopted longitudinal research designs. Cross-sectional studies usually collect data at a single point in time, which allows them to provide a snapshot of the current situation [58]. In contrast, longitudinal studies involve the long-term tracking of participants to identify any possible patterns of change over time [58]. For social science studies, such longitudinal data are of utmost importance as they enable the efficient and unbiased estimation of parameters in any dynamic process [59]. The limited number of longitudinal studies indicates a lack of research on topics such as the long-term washback effects of the test. For instance, Huang and Zeng conducted a longitudinal study and used mixed methods to examine the sustainability of the washback effects of the NMET [41]. They collected data from high school teachers three times over a five-year period. The results showed changes in the teachers' perceptions of the test and their attitudes toward the test reforms. Such changes were due to several factors, including the design of the test and pressure from the social environment, among others. In the long-term process, these factors may dynamically affect the washback of the test, which can only be explored through longitudinal research. Given the limited amount of longitudinal washback research, future researchers are encouraged to employ longitudinal research.

5.2. Research Foci of the Studies

Three main research themes were identified. The first major theme pertains to the washback effects on language learning[30,41,48]. For example, Zhang conducted a large-scale questionnaire survey to examine Chinese university students' perceptions of NMET[45]. The study revealed that the participants generally believed the NMET helped enhance their English learning outcomes. In Dong et al.'s study of Chinese high school students, the results showed that the NMET influenced their learning practices, English learning motivation, and anxiety[60]. The second major theme is about the washback on language teaching. For example, Zhang and Kong found that HSK had an impact on the language teachers' teaching content, teaching method, teaching goal, and teaching behaviors[30]. The test also influenced the formulation of school policies, curriculum arrangements, and textbook selection. These two research themes identified in this study echoed the findings of previous systematic reviews of washback research[35,36]. Paxton and his colleagues carried out a systematic review of 26 studies that examined the washback of the Japanese University English Entrance Exam[35]. They identified four major research themes: effects on learners' behaviors, effects on teachers' behaviors, effects on the psychology of teachers and students, and effects on students' English listening abilities. In a review of washback studies conducted in eight South Asian countries, the researcher found that language tests have a significant influence on language teaching and learning in these countries[36]. For instance, the test can have an impact on teachers' teaching methods, teaching activities, and the selection of learning materials, among others. In addition, the tests affected students' learning behaviors, perceptions of the test, and learning motivation, among other factors. These findings suggest that researchers have always been interested in how tests affect the two crucial aspects: language teaching and learning[23–26]. In the early stage of washback research, scholars primarily concentrated on exploring whether washback effects exist and, if they do, what the specific washback effects are[23–26]. For example, in Hughes's trichotomy model, three parts were influenced by language tests: participants (students, teachers, etc.), process (teaching and learning processes), and product (learning outcomes)[2]. Bailey proposed that the test can not only directly influence the test-takers, but also have an impact on other stakeholders relevant to language teaching, such as language teachers, school administrators.

Furthermore, the third theme concerns the factors that influence the washback effects. For instance, in Zhang's study, the participants' perceptions were influenced by the region, specifically the province where they took the test[45]. Dong et al. found that the washback effects varied significantly among students in terms of gender, grade, and English proficiency[60]. The test had a stronger influence on the English learning motivation of male students. Female students had significantly higher learning anxiety levels and spent more time on English learning. The third theme, which has long been discussed in prior washback research, reflects a consistent and longstanding academic interest in this domain[7,28]. As related research has progressed, it has become evident that the washback effect of language tests is more complex and could be affected by a multitude of factors[7,61,62]. Watanabe classified these factors into four aspects: personal factors, micro-environmental factors, macro-environmental factors, and test factors[62]. In Shi's model of washback, internal, external, and test-related factors together affect the washback effects on students' language learning and psychology[22]. Dong suggested that test-related, personal, and environmental factors are not isolated but rather interwoven and interact with one another[7]. They jointly influence the washback of a language test. Despite the different terms, these scholars' classifications consistently cover three core aspects: personal (e.g., students' and teachers' perceptions and attitudes), test-related (e.g., test design, importance, and difficulty), and social-contextual (e.g., institutional policies and cultural values). This tripartite structure highlights the fundamental interaction among individuals, tests, and the social structures in shaping washback mechanisms.

A model of washback was proposed based on previous theoretical models of washback and the findings of this study (See **Figure 3**). According to the model, a language test can influence multiple aspects of language learning and teaching. Specifically, it has an impact on language learning in terms of the process, learning outcomes, and learner psychology[30,41,48]. It also exerts effects on language teaching, including the teaching process and teacher psychology. Furthermore, this model includes three types of factors that can influence the washback effects: personal, test-related, and social context factors. This model integrates and synthesizes

previously fragmented elements of washback into a coherent framework and highlights the multi-layered nature of how language tests influence educational practices. Moreover, by incorporating the three types of factors, it offers a more nuanced understanding of why washback varies across different contexts (**Figure 5**).

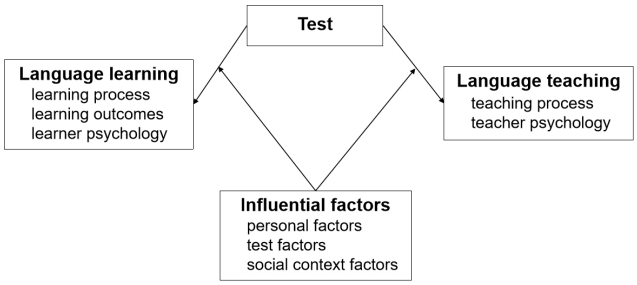

Figure 5. A proposed model of washback.

6. Conclusions

To summarize, this study reviewed washback research conducted in the Chinese mainland from 2005 to 2025. It presents a clear overview of the selected studies, including language tests, participants, research methodology, and research focus. Based on the research findings, this study offers suggestions for future research. Firstly, since the majority of the reviewed studies have centered around the washback of English tests, it is recommended that future scholars shift their focus and explore the washback of non-English tests, such as the HSK. So far, the number of studies on the washback of non-English language tests is relatively small. However, their research significance should not be underestimated. Given the distinct purposes and target populations of non-English language tests, these tests may generate unique washback effects. Therefore, further research is necessary on the washback effects of non-English language tests. Additionally, as shown in this study, over 80% of the reviewed studies involved university students and teachers as participants. Therefore, future scholars should consider recruiting diverse participants, including students and teachers from high schools, middle schools, and vocational schools, as well as other relevant stakeholders, such as school leaders and test designers. Such studies can offer more meaningful implications for stakeholders at various educational levels. In addition, as most of the reviewed studies used quantitative or mixed methods and were cross-sectional, it is recommended that future researchers employ more qualitative methods and longitudinal designs. This can provide in-depth insights into the dynamic changes and underlying mechanisms of washback effects over time. Furthermore, among the reviewed studies, only one study compared the washback of different tests [9], while others all examined the washback of a single test. Different tests may have different washback effects, and the factors that affect these effects may also vary. In light of this, future scholars can compare the influences of different language tests on language learning and teaching and explore whether there are any similarities or differences. This will enrich the existing knowledge of washback research. Moreover, this study proposed a model of washback effects on language learning, which could serve as a comprehensive framework for understanding the complex washback mechanisms. Future scholars can utilize this model to carry out more in-depth empirical research and to further refine it.

This review has several limitations that need to be acknowledged. Firstly, this review only involved washback research conducted in the Chinese mainland, whereas studies carried out in other countries or regions were not included. Also, the literature search was restricted to only three databases. Only peer-reviewed journal articles published between 2005 and 2025 were included, while other relevant publications, such as book chapters, theses, and conference papers, were excluded. The narrow scope of literature sources may affect the comprehensiveness of the research findings and lead to an incomplete understanding of the topic. Another limitation is the lack of quality appraisal tools during both the literature screening process and the review of the methodology of the included studies. Although the reviewers attempted to ensure the reliability of the literature selection and evaluation of the studies, the judgment of research quality could be subjective without the use of formal quality assessment tools. This could undermine the objectivity of the evaluation and the credibility of the review.

Funding

This work received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare that they have no conflict of interest.

References

[1] Alderson, J.C., Wall, D., 1993. Does washback exist? Applied Linguistics. 14(2), 115–129. DOI: https://doi.org/10.1093/applin/14.2.115

[2] Hughes, A., 1989. Testing for Language Teachers, 1st ed. Cambridge University Press: Cambridge, UK.

[3] Baker, E., 1991. Alternative assessment and national policy. In Proceedings of the National Research Symposium on Limited English Proficient Students' Issues: Focus on Evaluation and Measurement, Washington, DC, USA, 1991.

[4] Messick, S., 1989. Validity. In: Linn, R.L. (ed.). Educational Measurement, 3rd ed. American Council on Education/Macmillan: New York, NY, USA. pp. 13–103.

[5] Green, A., 2020. Washback in language assessment. The Encyclopedia of Applied Linguistics. 1–6. DOI: https://doi.org/10.1002/9781405198431.wbeal1274.pub2

[6] Xie, Q., Jia, Q., 2025. Three decades of research on washback (1993–2023): A bibliometric study. Language Testing in Asia. 15(1), 1–29. DOI: https://doi.org/10.1186/s40468-025-00357-w

[7] Dong, M., 2019. Some basic issues needed to clarify in washback studies. Foreign Language Learning Theory and Practice. (3), 50–57. (in Chinese)

[8] Chen, D., 2022. A review of empirical studies on washback of English language testing in China. Creative Education Studies. 10(8), 1860–1866. DOI: https://doi.org/10.12677/ces.2022.108294

[9] Xiao, W., Gu, X.D., 2022. A comparative study on the washback mechanisms of CET-6, IELTS and TOEFL writing modules. Foreign Language Learning Theory and Practice. 3, 94–103.

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

[11] Zhang, J., 2021. A moderated mediation analysis of the relationship between a high-stakes English test and test takers' extracurricular English learning activities. Language Testing in Asia. 11(5). DOI: https://doi.org/10.1186/s40468-021-00120-x

[12] Li, H., Zhong, Q., Suen, H.K., 2012. Students' perceptions of the impact of the College English Test. Language Testing in Asia. 2(3). DOI: https://doi.org/10.1186/2229-0443-2-3-77

[13] Cheng, Y., Hamid, M.O., 2025. Social impact of Gaokao in China: A critical review of research. Language Testing in Asia. 15(1). DOI: https://doi.org/10.1186/s40468-025-00355-y

[14] Zhang, X., 2021. Stakeholders' test perceptions on test reform. Studies in Educational Evaluation. 70, 101064. DOI: https://doi.org/10.1016/j.stueduc.2021.101064

[15] Hughes, A., 1993. Backwash and TOEFL 2000 [Unpublished manuscript]. University of Reading: Reading, UK.

[16] Bailey, K.M., 1996. Working for washback: A review of the washback concept in language testing. Language Testing. 13(3), 257–279. DOI: https://doi.org/10.1177/026553229601300303

[17] Green, A., 2006. Watching for Washback: Observing the Influence of the International English Language Testing System Academic Writing Test in the classroom. Language Assessment Quarterly. 3(4), 333–368.

[18] Wen, X., Chano, J., 2024. A critical review on washback effect in education and its influence on curriculum design. Forum for Linguistic Studies. 7(1), 287–297. DOI: https://doi.org/10.30564/fls.v7i1.7893

[19] McNamara, T., 2006. Language testing: The social dimension. International Journal of Applied Linguistics. 16(2), 242–258. DOI: https://doi.org/10.1111/j.1473-4192.2006.00117.x

[20] Green, A., 2007. IELTS Washback in Context: Preparation for Academic Writing in Higher Education. Cambridge University Press: Cambridge, UK. pp. 1–31.

[21] Qi, L., 2012. Washback studies revisited. Modern Foreign Languages. 35(2), 202–208, 220. (in Chinese)

[22] Shih, C., 2007. A new washback model of students' learning. The Canadian Modern Language Review. 64, 135–162. DOI: https://doi.org/10.3138/cmlr.64.1.135

[23] Qi, L., 2004. The intended washback effect of the National Matriculation English Test in China: Intentions and reality. Foreign Language Teaching and Research. 36(5), 357–363.

[24] Alderson, J.C., Hamp-Lyons, L., 1996. TOEFL preparation courses: A study of washback. Language Testing. 13(3), 280–297. DOI: https://doi.org/10.1177/026553229601300304

[25] Gu, X., 2007. An empirical study of CET washback on college English teaching and learning in China. Journal of Chongqing University (Social Science Edition). 13(4), 119–125.

[26] Hayes, B., Read, J., 2004. IELTS test preparation in

New Zealand: Preparing students for the IELTS academic module. In: Cheng, L., Watanabe, Y., Curtis, A. (eds.). Washback in Language Testing: Research Context and Methods. Lawrence Erlbaum Associates: Mahwah, NJ, USA. pp. 97–112.

[27] Green, A., 2006. Washback to the learner: Learner and teacher perspectives on IELTS preparation course expectations and outcomes. Assessing Writing. 11(2), 113–134. DOI: https://doi.org/10.1016/j.asw.2006.07.002

[28] Li, Y., Hu, B., 2025. Unpacking the washback of the HSK on international Chinese language education: A study of the influencing factors. Language Testing in Asia. 15(1). DOI: https://doi.org/10.1186/s40468-025-00370-z

[29] Kong, F., Zhang, Y., 2024. Investigating the washback of the Chinese Language Proficiency Test (HSK): From the perspective of CSL students. Sage Open. 14(1). DOI: https://doi.org/10.1177/21582440231224599

[30] Zhang, Y.L., Kong, F.Y., 2021. Washback of high-stake language tests: Implications for the integration of testing, teaching and learning—Take the HSK as an example. Technology Enhanced Foreign Language Education. 3, 76–82, 108, 12. DOI: https://doi.org/10.20139/j.issn.1001-5795.2021.03.010 (in Chinese)

[31] Liu, J., 2025. Washback effects of the Portuguese CAPLE exams from Chinese university students and teachers' perspectives: A mixed-methods study. Language Learning in Higher Education. 15(1), 221–243. DOI: https://doi.org/10.1515/cercles-2024-0038

[32] Jia, F., Kong, Y., Yi, H., 2022. Washback of a school-based College English Proficiency Test on science and engineering majors. Foreign Language World. (2), 56–62.

[33] Li, X., 1990. How powerful can a language test be? The MET in China? Journal of Multilingual and Multicultural Development. 11(5), 393–404. DOI: https://doi.org/10.1080/01434632.1990.9994425

[34] Shohamy, E., Donitsa-Schmidt, S., Ferman, I., 1996. Test impact revisited: Washback effect over time. Language Testing. 13(3), 298–317. DOI: https://doi.org/10.1177/026553229601300305

[35] Paxton, S., Yamazaki, T., Kunert, H., 2022. Japanese university English language entrance exams and the washback effect: A systematic review of the research. Journal of Pan-Pacific Association of Applied Linguistics. 26(2), 1–20. DOI: https://doi.org/10.25256/paal.26.2.1

[36] Allen, D., Tahara, T., 2021. A review of washback research in Japan. JLTA Journal. 24, 3–22. DOI: https://doi.org/10.20622/jltajournal.24.0_3

[37] Sultana, N., 2018. A brief review of washback studies in the South Asian countries. The Educational Review, USA. 2(9), 468–474. DOI: https://doi.org/10.26855/er.2018.09.002

[38] Page, M.J., Moher, D., Bossuyt, P.M., et al., 2021. PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. BMJ. 372, 1–33. DOI: https://doi.org/10.1136/bmj.n160

[39] Zhang, H., Bournot-Trites, M., 2021. The long-term washback effects of the National Matriculation English Test on college English learning in China: Tertiary student perspectives. Studies in Educational Evaluation. 68, 100977. DOI: https://doi.org/10.1016/j.stueduc.2021.100977

[40] Liu, X., Yu, J., 2021. Relationships between learning motivations and practices as influenced by a high-stakes language test: The mechanism of washback on learning. Studies in Educational Evaluation. 68, 100967. DOI: https://doi.org/10.1016/j.stueduc.2020.100967

[41] Huang, L., Zeng, J., 2025. Sustainability of washback effects: A longitudinal study of China's twice-yearly NMET reform. Language Testing in Asia. 15(1). DOI: https://doi.org/10.1186/s40468-025-00358-9

[42] Wang, J., Zheng, Y., Zou, Y., 2024. Face validity and washback effects of the shortened PTE Academic: Insights from teachers in Mainland China. Language Testing in Asia. 14(1). DOI: https://doi.org/10.1186/s40468-024-00302-3

[43] Gu, X.D., Yang, Z.Q., Liu, X.H., 2013. A longitudinal study of the washback effect of CET on college English classroom teaching: Revisiting the classrooms of three college English teachers. Foreign Language Testing and Teaching. 1, 18–29, 63. (in Chinese)

[44] Guo, S.H., Li, F.X., 2012. Washback effect of Internet-based College English Test on college English teachers' professional development. Technology Enhanced Foreign Language Education. 5, 72–76. (in Chinese)

[45] Zhang, H., 2024. The washback of the National Matriculation English Test on senior high school English learning outcomes: Do test takers from different provinces think alike? Language Testing in Asia. 14(1). DOI: https://doi.org/10.1186/s40468-024-00286-0

[46] Zhang, H., Zhang, W.X., 2020. English teachers' perceptions of NMET washback on senior high school English teaching and learning—Based on a large-scale nationwide survey. Foreign Language Learning Theory and Practice. 3, 36–45.

[47] Kong, F.Y., Zhang, Y.L., 2021. A Research of HSK test's washback effect on teaching Chinese as a second language. Journal of Tianjin Normal University (Social Sciences). 4, 46–51. (in Chinese)

[48] Gong, K., 2023. Challenges and opportunities for spoken English learning and instruction brought by automated speech scoring in large-scale speaking tests: A mixed-method investigation into the washback of SpeechRater. Asian-Pacific Journal of Second and Foreign Language Education. 8(1), 1–23. DOI: https:

//doi.org/10.1186/s40862-023-00197-2

[49] Wang, L., Wagner, E., 2020. Washback of the College English Test-Band Four on English teaching and learning in China. The Journal of Asia TEFL. 17(4), 1214–1235. DOI: https://doi.org/10.18823/asiatefl.2020.17.4.4.1214

[50] Chen, Q., Hao, C., Xiao, Y., 2020. When testing stakes are no longer high: Impact on the Chinese College English learners and their learning. Language Testing in Asia. 10(1). DOI: https://doi.org/10.1186/s40468-020-00102-5

[51] Fan, J., Frost, K., Liu, B., 2020. Teachers' involvement in high-stakes language assessment reforms: The case of Test for English Majors (TEM) in China. Studies in Educational Evaluation. 66, 100898. DOI: https://doi.org/10.1016/j.stueduc.2020.100898

[52] Wang, J.Y., Sun, Y., 2018. A study on the washback effect of the Japanese Language Proficiency Test (JLPT). Foreign Languages Bimonthly. 41(6), 73–80, 89. (in Chinese)

[53] Dunifa, L., 2023. Evaluating oral English program for non-English major students: Focusing on self-assessment of students' speaking abilities and their needs. Novitas-ROYAL (Research on Youth and Language). 17(2), 34–49. DOI: https://doi.org/10.5281/zenodo.10015757

[54] Yang, Y., Jirawit, Y., 2025. Navigating HSK Level 5 challenges: Dual perspectives from Thai undergraduates and Chinese lecturers. Language Testing in Asia. 15(1). DOI: https://doi.org/10.1186/s40468-025-00360-1

[55] Laotrakunchai, K., Laotrakunchai, Y., Chuwicharoenkit, K., 2021. The backwash effect of the new HSK Level 4 test on high school students in Thailand. Journal of Genomics. 15(1), 142–168.

[56] Litosseliti, L., 2010. Research Methods in Linguistics. Continuum: London, UK. pp. 49–67.

[57] Creswell, J.W., Creswell, J.D., 2014. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Sage: Thousand Oaks, CA, USA. pp. 1–24.

[58] Dörnyei, Z., 2007. Research Methods in Applied Linguistics: Quantitative, Qualitative and Mixed Methodologies. Oxford University Press: Oxford, UK. pp. 78–92

[59] Menard, S., 2007. Handbook of Longitudinal Research Design, Measurement, and Analysis. Academic Press: New York, NY, USA. pp. 1–12

[60] Dong, M., Fan, J., Xu, J., 2023. Differential washback effects of a high-stakes test on students' English learning process: Evidence from a large-scale stratified survey in China. Asia Pacific Journal of Education. 43(1), 252–269. DOI: https://doi.org/10.1080/02188791.2021.1918057

[61] Sakarya Akbulut, H., Mirici, İ.H., 2024. An investigation into EFL students' perceptions of task-based language assessment in the blended learning environment. Novitas-ROYAL (Research on Youth and Language). 18(2), 12–28. DOI: https://doi.org/10.5281/zenodo.13225030

[62] Watanabe, Y., 2004. Methodology in washback studies. In: Cheng, L., Watanabe, Y., Curtis, A. (eds.). Washback in Language Testing: Research Contexts and Methods. Lawrence Erlbaum Associates: Mahwah, NJ, USA. pp. 19–36.