Malaysian Journal of Social Sciences and Humanities (MJSSH) (e-ISSN : 2504-8562)
2025, Volume 10, Issue 12, e003682
DOI: https://doi.org/10.47405/mjssh.v10i12.3682

RESEARCH ARTICLE

The Washback of the IELTS Speaking Test on Chinese Test-Takers' Perceptions and Preparation Behaviours

Liu Yuan[1*], Alla Khan[2]

[1]School of Languages, Literacies and Translation, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia
Email: yuan.liu7@outlook.com
[2]School of Languages, Literacies and Translation, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia
Email: allabaksh@usm.my

CORRESPONDING AUTHOR (*):
Liu Yuan
(yuan.liu7@outlook.com)

KEYWORDS:
IELTS speaking test
Washback
Mediating factors
Test Preparation course
Test score

CITATION:
Liu, Y., & Alla, K. (2025). The Washback of the IELTS Speaking Test on Chinese Test-Takers' Perceptions and Preparation Behaviours. *Malaysian Journal of Social Sciences and Humanities (MJSSH), 10*(12), e003682. https://doi.org/10.47405/mjssh.v10i12.3682

ABSTRACT

In the Chinese EFL context, a critical paradox exists: despite widespread enrolment in IELTS preparation courses, test-takers continue to struggle to achieve satisfactory speaking scores. This study investigates this issue through the lens of the washback effect—the influence of tests on learning. Drawing on a mixed-methods questionnaire completed by 236 Chinese IELTS test-takers, this study employed descriptive statistics, structural equation modelling (SEM), and thematic analysis. The quantitative results revealed a significant misalignment between test-takers' interpretations of the test design and the official assessment criteria. While the SEM indicated no significant direct relationship between key mediating factors (self-perceived proficiency, academic expectations, and individual differences) and test scores, qualitative findings offered a crucial explanation. Specifically, many learners' misinterpretations led them to rely on rote memorisation rather than developing the communicative competence the test aims to measure. This study extends washback research by demonstrating that mediating factors shape preparation behaviours through indirect pathways (processes) rather than directly determining outcomes (products). The findings underscore the need for test developers and educators to bridge the gap between test-takers' beliefs and assessment objectives to promote effective learning.

Contribution/Originality: This study contributes to the existing literature by extending the mechanism of washback to include mediating factors. It provides empirical evidence that test-takers' perceptions and individual differences shape preparation behaviors through indirect pathways, offering a new explanation for the "high enrollment, low achievement" paradox in the Chinese IELTS context.

1. Introduction

The impact of testing on learners and learning, commonly referred to as washback, has received growing scholarly attention in recent years (Yu et al., 2017; Sato, 2019). Early studies conceptualised washback as a direct influence of testing on what and how learners study (Alderson & Wall, 1993; Messick, 1996). Subsequent research, however, has revealed that the relationship between testing and learning is far from linear; rather, it is a complex and dynamic phenomenon mediated by various contextual and individual factors (Green, 2007; Xie, 2013). Recent studies have explored how test-related and non-test-related factors interact to shape learning processes (Yu et al., 2017; Sato, 2018; Dong, 2020). Yet, limited research has examined how test-takers' perceptions of test design influence their preparation behaviours and subsequent performance.

This study seeks to address this gap by investigating how a high-stakes test can influence test-takers' learning in the Chinese EFL context, specifically focusing on their perceptions of the IELTS Speaking Test design and the learning behaviours adopted in preparation for it. The study focuses on IELTS speaking preparation courses, which are a common learning pathway among Chinese students seeking to meet university language entry requirements (Hu & Trenkic, 2021; Ma & Chong, 2022). Despite the popularity of such courses, Chinese IELTS candidates' average speaking band score (5.55) remains below many universities' required threshold (IELTS, 2019; IELTS, 2021). This contradiction between high enrolment and low achievement necessitates an investigation into how learners' interpretations of the test connect with their preparation strategies and subsequent scores.

Early washback studies emphasised direct test effects on learning (Alderson & Wall, 1993), whereas later research acknowledged the importance of mediating factors that shape the indirect effects of tests (Xie, 2013; Sato, 2019; Dong, 2020). Nonetheless, few studies have examined how test-takers' interpretation of test design and their reasons for enrolling in preparation courses relate to their test results, particularly within the Chinese EFL context. By exploring the mediating roles of self-perceived proficiency, academic expectations, and individual differences, this study provides empirical evidence of the indirect and multifaceted pathways through which washback operates. This highlights how test-takers' interpretation of the target test, reasons for undertaking a test preparation course and test scores, offering fresh insights into the mechanism through which washback manifests in test preparation contexts.

2. Literature review

2.1. Washback: test-takers' perceptions of the test design

In this study, *test design* refers to the structure, assessment criteria, and skill focus of the IELTS Speaking Test, such as task types and scoring rubrics (IELTS, 2020). Understanding test-takers' perceptions of the test design is crucial in washback research because learners' preparation behaviours are strongly influenced by how they interpret the test's evaluation goals (Hughes, 1993; Bailey, 1996; Xie, 2011).

Hughes (1993) expanded on Alderson and Wall's (1993) washback hypotheses to create a framework that included participants, processes, and products, arguing that the perceptions and attitudes of all participants, including test-takers, are susceptible to the influence of tests. Building upon this, Bailey (1996) proposed a foundational model of

washback, stressing that test users must accurately understand the test's purpose, the constructs it measures, and the uses of its results.

In recent years, there has been increasing scholarly interest in test-takers' perceptions (Wei, 2017; Booth, 2018; Ma, 2017), driven by the recognition that washback is not produced solely by the test itself but also by how test-takers interpret it. Chalhoub-Deville and O'Sullivan (2020) further argued that examining test-takers' perceptions contributed to a deeper understanding of test consequences and their relationship to validity. Misalignment between test-takers' perceptions and the intentions of test developers can lead to the inappropriate use of otherwise well-designed tests (Popham, 1997). Kane (2013) emphasised that consequences are a core aspect of validity, and that students—key stakeholders in the testing process—experience these consequences, both intended and unintended, through their learning processes and outcomes. Accordingly, exploring test-takers' perceptions is indispensable for a comprehensive understanding of washback and for evaluating the consequential validity of language assessments.

2.2. Washback on learning and Mediating factors

Research on washback has increasingly emphasized that the effects of high-stakes testing extend beyond the test content itself. While earlier studies often conceptualised washback as a direct and linear relationship between a test and learners' behaviours, more recent work has highlighted its complexity. Washback is now understood as a multifaceted phenomenon, mediated by a range of individual and contextual factors that interact with one another to influence both learning processes and outcomes (Green, 2007; Xie & Andrews, 2013). In this study, test scores are considered outcomes rather than processes, while the primary focus is on how enrolment reasons and mediating factors shape both learning behaviours and performance.

Following the discussion of including test-takers' perception of the target test design into washback studies, some recent studies have found that the washback of a test on test-takers' learning is not only limited to the content of the test, but also to some other mediating factors (Green, 2007; Xie & Andrews, 2013). This suggests that washback should no longer be conceptualised as a single, direct effect of the test on learners. Instead, the washback effect is considered to be a complex phenomenon in which different factors might intervene with each other and then have an impact on learning behaviours and learning outcomes, namely test scores. In this study, test scores are outcomes, not learning processes. This study explores how enrolment reasons and mediating factors influence both processes and outcomes.

Many earlier studies framed washback as the test exerting a direct influence on learners' behaviours, such as the activities they engaged in or the skills they prioritised in response to their interpretations of the test construct (Mizutani, 2009; Chappell et al., 2019; Zhang & Bournot-Trites, 2021). For example, Green's (2007) seminal study on IELTS preparation courses in the UK found that participation in such courses did not necessarily guarantee score improvement. Instead, score gains were most evident among students who planned to retake the test and who had previously achieved low writing scores. Green (2007) concluded that learning outcomes were driven less by the course itself and more by learners' personal goals and their interpretations of test demands. His findings underscored the importance of moving beyond a causal model of

washback to one that accounts for the mediating influence of external and internal factors.

Building on this perspective, subsequent studies have confirmed that test-takers' learning behaviours and outcomes are not shaped solely by the test but also by other mediating influences, some of which are construct-irrelevant (Cheng et al., 2015; Zhan & Wan, 2016; Yu et al., 2017; Sato, 2019). These mediating factors help explain why learners enrolled in the same preparation course may experience different outcomes (Gosa, 2004; Cheng & Deluca, 2011; Dong, 2020; Tsang & Isaacs, 2022). A review of the literature reveals three mediating factors that are most consistently linked to washback on learning.

The first mediating factor was self-perceived language proficiency. This mediator refers to test-takers' subjective evaluation of their speaking ability to meet test requirements (Cheng & Deluca, 2011; Fox & Cheng, 2007; Gan, 2009). The second factor is the academic expectations to take the test preparation course (Gosa, 2004; Green, 2007; Xie & Andrews, 2013). This refers to learners' goals for score improvement or skill development through prep courses (Gosa, 2004). The final one is individual differences (Ferman, 2004; Horwitz, 2010; Shih, 2007; Mickan & Motteram, 2009). This mediator could be seen as how test-takers' learning activities could be affected by personal factors, including their anxiety level about the test and taking advice from their peers or parents. Therefore, the current study aims to include all these three commonly referred mediating factors in the existing literature to explore how the interplay between these factors could have an impact on Chinese EFL IELTS test-takers' learning processes, namely reasons for undertaking a test preparation course and then test scores.

In conclusion, the literature has shifted from a linear perspective on washback to a complicated model that includes mediating factors. However, a critical gap remains in understanding how these factors—self-perceived proficiency, academic expectations, and individual differences—collectively interact within a specific learning context to influence both the decision to prepare (process) and the ultimate test scores (product). The test preparation course, as a widespread and high-stakes context, provides an ideal site to observe this complex interplay in action.

2.3. Test preparation course context

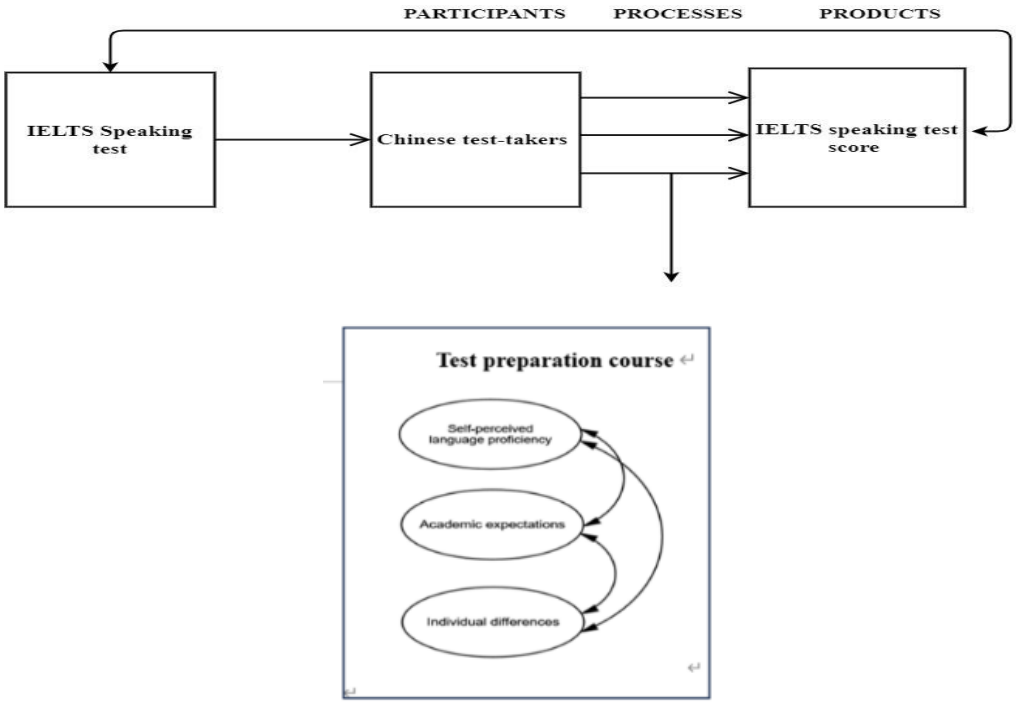
Test preparation courses, often referred to as *coaching*, are short-term instructional programmes designed to familiarise learners with the structure and demands of a specific test, distinct from general language instruction (Messick, 1981). Research on such courses for high-stakes assessments like IELTS has expanded considerably in recent years (He et al., 2024). While studies in English-speaking contexts report mixed results regarding their effectiveness in improving scores (Green, 2007; Hayes & Read, 2004; Hawkey, 2006), research in EFL contexts remains relatively scarce. Recent evidence suggests that test-takers' enrolment decisions are influenced by a complex combination of factors extending beyond the test itself (Yu et al., 2017; Hu & Trenkic, 2021). However, there is still limited understanding of how different mediators interact to shape both these decisions and performance outcomes. The present study focuses on the IELTS Speaking Test in China to investigate these relationships systematically.

2.4. The framework for researching washback on learning

The conceptual framework adopted for this study is an adaptation of Bailey's (1996) model of washback, which posits that washback operates through interactions among *Participants* (test-takers), *Processes* (learning and preparation behaviours), and *Products* (test outcomes). Although Bailey's (1996) model has been critiqued for lacking empirical validation, its tripartite structure aligns closely with this study's research focus—that learning behaviours are mediated by a range of factors that, in turn, may influence test performance.

This framework hypothesises that test-takers' perceptions of the IELTS Speaking Test design interact with mediating factors (V1–V3) to influence their preparation behaviours and eventual scores. As Hamp-Lyons (1997) and Wall (1997) noted, enriching the *Processes* component of Bailey's (1996) model with empirical data contributes to a more nuanced understanding of how washback operates. Consequently, the current study extends the model by empirically exploring how perceptions, mediators, and outcomes are connected in a Chinese EFL context. Therefore, a conceptual framework was developed by adapting Bailey's (1996) model (Figure 1). In this framework, arrows represent hypothesised influences of test design, interpretation and mediating factors (V1-V3) on preparation behaviours and scores, based on Bailey (1996).

Figure 1: Conceptual Framework



2.5. Research Questions

The following research questions were proposed to examine the proposed relationship raised in the literature review:

i. What is the relationship between Chinese IELTS test-takers' interpretation of the IELTS speaking test design and the perceptions of the intended washback?

ii. How could the interplay between the mediating factors have an impact on the intended washback of the IELTS speaking test on test-takers' learning behaviours and test score?

3. Methodology

3.1. Research Design

This study employed a convergent parallel mixed-methods design, utilising a single self-administered questionnaire to collect quantitative and qualitative data simultaneously. The instrument comprised a quantitative section with closed-ended multiple-choice and Likert-scale questions to quantify participants' perceptions, and a qualitative section with open-ended prompts to elicit detailed explanatory feedback. The quantitative data were analysed using descriptive statistics, Spearman correlation, and Structural Equation Modelling (SEM), while the qualitative data underwent inductive thematic analysis. The two datasets were then integrated during the interpretation phase, allowing for the triangulation of findings to converge upon a more complete and nuanced understanding of the research problem.

3.2. Instrument

A questionnaire was adapted from previous studies (Yu et al., 2017; Tsang, 2017) to answer the research questions in this study. The questionnaire was designed to collect both closed-ended and open-ended responses, thereby combining quantitative and qualitative data within a single instrument. This integration allowed for a more holistic understanding of the washback phenomenon, enabling cross-validation between data types (Creswell & Plano Clark, 2017).

The questionnaire consisted of three parts. Part one gathered demographic information, including participants' gender, the preparation course they attended, and their IELTS Speaking Test scores. Part Two explored test-takers' interpretations of the IELTS Speaking Test design, focusing on their understanding of the skills being assessed and their self-perceived evaluation goals. Multiple-choice items were used to assess comprehension of the test's design features, while 5-point Likert-scale items elicited participants' self-perceptions of test criteria and assessment goals. Each quantitative item was followed by an open-ended question to allow participants to elaborate on their reasoning and opinions. Part three examined test-takers' reasons for enrolling in an IELTS preparation course. This section included both Likert-scale and open-ended questions designed to capture the range of motivational and contextual factors influencing enrolment.

Prior to the main study, the questionnaire's validity and reliability were established through a pilot study involving 150 IELTS test-takers. For content validity, five experts—each with over ten years of IELTS teaching and assessment experience—reviewed the instrument to ensure that the items aligned with IELTS rubrics and that the questions were clear and relevant. For construct validity, Principal Component Analysis (PCA) was conducted using *JAMOVI* (version 2.3.26). The Kaiser–Meyer–Olkin (KMO) measure was 0.864, and Bartlett's test of sphericity was significant ($p < 0.01$), indicating that the data were suitable for factor analysis.

Regarding reliability, Cronbach's alpha coefficients were computed for each subscale and for the overall questionnaire. The subscale assessing test-takers' interpretation of the IELTS Speaking Test design (multiple-choice and Likert-scale items) had an alpha of 0.70, indicating acceptable reliability. The subscale assessing reasons for enrolling in preparation courses yielded an alpha of 0.858, indicating good reliability. The overall scale achieved an alpha of 0.814, suggesting satisfactory internal consistency (Larson-Hall, 2015). On this basis, the instrument was deemed both valid and reliable for use in the main study.

3.3. Sampling and Participants

A convenience sampling method was used to recruit participants. While this non-probability sampling technique limits the generalizability of the findings, it was deemed the most feasible and practical approach for accessing a geographically dispersed and specific population of recent IELTS test-takers in China. Individuals were sourced from IELTS preparation centers across various cities and online platforms, with some participants completing the course remotely for test preparation. The final sample consisted of N = 236 participants. The sample size was determined based on the requirements for Structural Equation Modelling (SEM). Following the recommendation of Kline (2016), a sample size of over 200 is considered adequate for most SEM analyses. This target was pursued to ensure sufficient statistical power for the planned model testing.

3.4 Data Analysis

Quantitative data were analysed using a combination of statistical software: JAMOVI (version 2.3.26) for descriptive statistics, SPSS (version 27) for Spearman correlation analysis, and AMOS (version 24) for confirmatory factor analysis (CFA).

Descriptive statistics were used to summarise participants' interpretations of the test design and perceptions of the intended washback. Spearman correlation analysis was conducted to examine relationships among variables, given that normality assumptions were not met. SEM (Structural equation modelling) was employed to test the proposed model of the relationships between mediating factors, preparation behaviours, and test scores. Qualitative data from open-ended responses were analysed using an inductive thematic analysis approach, enabling patterns and themes to emerge from participants' perspectives.

3.5. Ethical Consideration

This study obtained ethical approval from the Universiti Sains Malaysia (USM) Human Research Ethics Committee (JEPeM). All procedures performed in this study involving human participants were conducted in accordance with the ethical standards of this institutional research committee and with the 1964 Helsinki declaration and its later amendments. Informed consent was obtained from all individual participants included in the study.

4. Results

As stated in the previous part, the quantitative results will be shown in two sections, including the descriptive statistics and correlation analysis for RQ1 and CFA outputs for RQ2. In addition, a qualitative analysis of the open-ended responses was conducted.

4.1 Descriptive Statistics

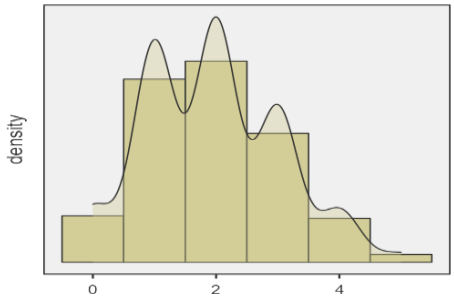
4.1.1. RQ1: What is the relationship between Chinese IELTS test-takers' interpretation of the IELTS speaking test design and the perception of the intended washback?

First, Table 1 shows the descriptive statistics of Chinese IELTS test-takers' interpretation of the test design of the speaking test and Figure 2 shows the distribution of the data. The descriptive results show that test-takers achieved a mean score of 1.94 (SD = 1.10) out of 5 for their interpretation of the test design, suggesting limited understanding of the skills targeted by the IELTS Speaking Test. This indicates a partial misalignment between their perceptions and the test developers' intentions.

Table 1: Descriptive Statistics of test-takers' interpretation of test design

	N	Mean	SD
Interpretation of the test design	236	1.94	1.1
Self-perceived test criteria	236	21.8	2.25
Self-perceived skills	236	22.1	2.19

Figure 2: Test-takers' interpretation of test design



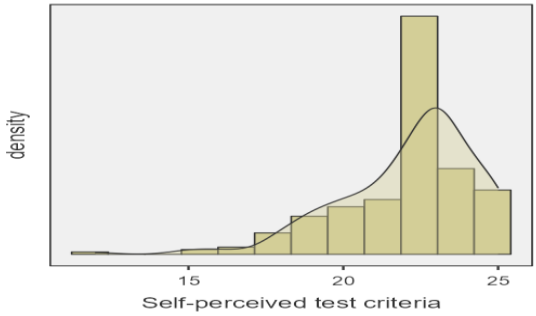
By contrast, participants' self-perceived understanding of test criteria (M = 21.8/25) and perceived skills (M = 22.1/25) were both high, suggesting that many learners believed they had a strong grasp of the test's evaluation goals. However, this self-confidence was inconsistent with their actual understanding of the test construct, as revealed by the design interpretation scores.

Then, the researchers employed a Likert scale to collect test-takers' perceptions of the intended washback, namely their self-perceived evaluation goals, including self-perceived skills and test criteria (Table 1). In the literature, the intended washback is defined as the impact of the test that the developers aim to have on the lives of test users (Cheng, 2005; Kane, 2013). For the current study, the intended washback refers to the influences of the IELTS speaking test on test-takers' interpretation of the test design and reasons for undertaking a test preparation course, and test score. Thus, the current study asked test-takers do self-reporting of their beliefs of skills the test aims to assess, and the test criteria will use in the real test as evidence to show to what extent test-

takers have perceived the evaluation goals of the target test. The researchers calculated the item responses of each Likert scale answer regarding their perceptions of skills (5 items/ 25 in full score) and test criteria (5 items/ 25 in total). According to the data presented in Figure 3 and Figure 4, the perception of test skills was found to be Mean=22.1/25, while the perception of test criteria was Mean=21.8/25.

Figure 3: Self-perceived test criteria

Figure 4: Self-perceived skills of the test



After illustrating the data from two types of items separately, the data comparison was made, and it showed a significant difference in how test-takers interpreted the test design and perceived its intended evaluation goals. To examine relationships between interpretation of the test design, perceived test criteria, and perceived skills, a Spearman correlation analysis was conducted due to the non-normal distribution of data. Normality testing using the Shapiro–Wilk test confirmed significant deviations from normality across all variables:

$W_1(236) = 0.919$, $p < 0.01$; $W_2(236) = 0.908$, $p < 0.01$; $W_3(236) = 0.906$, $p < 0.01$. Spearman's results indicated no significant correlations between test-takers' interpretation of the test design and either self-perceived skills ($r_s = 0.018$, $p = 0.778$) or self-perceived test criteria ($r_s = 0.076$, $p = 0.242$). This non-significant relationship suggests that many test-takers hold confident yet inaccurate beliefs about the test's construct, possibly leading them to adopt preparation strategies misaligned with the intended communicative goals of the IELTS Speaking Test.

4.2. SEM outputs

4.2.1. RQ2: How could the interplay between the mediating factors have an impact on the intended washback of the IELTS speaking test on test-takers' learning behaviours and test score?

In order to answer this research question, structural equation modelling (SEM) is conducted based on the responses from the questionnaire and the proposed SEM model of this study (Figure 5). The current study aims to explore how mediating factors could play a role in test-takers' learning behaviours, namely reasons to enrol in a test preparation course and their IELTS speaking score. Table 2 lists the name of different factors and observed variables and Table 3 displays the descriptive statistics of different factors.

Figure 5: Proposal SEM Model of this study

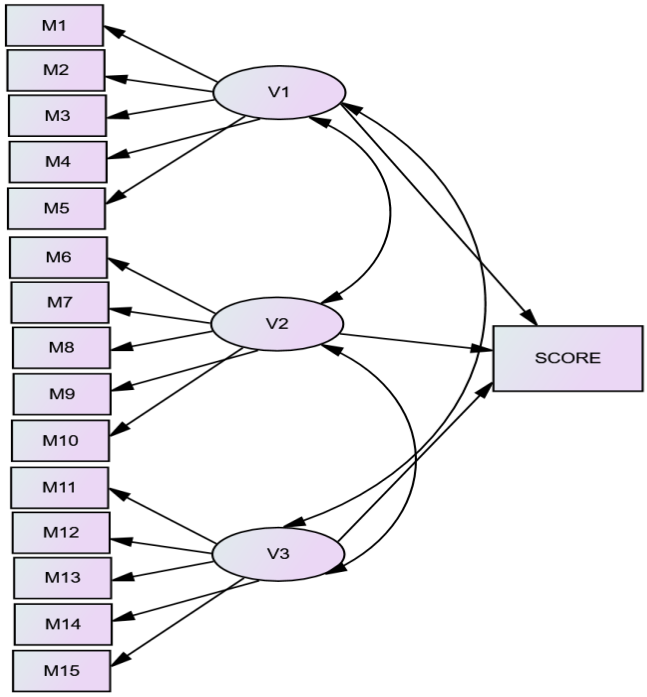


Table 2: Factors and observed variables

Code	Factor	Observed variable
M1	V1 Self-perceived Language Proficiency	My spoken English is not good enough to get the score I want
M2		I need to improve my speaking ability
M3		I need teachers to correct my pronunciation and intonation
M4		I need teachers to correct my grammar mistakes
M5		I could improve my speaking skill by learning from my classmates
M6	V2 Academic expectation	I want to improve my general speaking skills
M7		I want to learn test-taking strategies about the speaking test.
M8		I want to make myself familiar with the speaking test formats
M9		I want to learn some potential topics for my speaking test
M10		I want to get the score that my dream university required
M11	V3 Individual differences	it can help me reduce my test anxiety and the fear of the speaking test
M12		it helps me gain confidence in taking the speaking

		test
M13		I could study with my friends or classmates who also attend the course
M14		it keeps my parents happy
M15		I take advice from my teacher or mentor, because he/she thinks taking a course could help me to pass the test.
RTS	Recent test score	Recent test score

Table 3: Descriptive Statistics of SEM output

	N	Mean	Min	Max	Skewness	Kurtosis
M1	236	3.39	0	5	-0.484	-0.420
M2	236	4.02	0	5	-1.01	0.641
M3	236	3.58	0	5	-0.635	-0.255
M4	236	3.46	0	5	-0.468	-0.570
M5	236	3.60	0	5	-0.929	0.476
M6	236	4.01	0	5	-1.24	2.19
M7	236	4.03	0	5	-1.19	0.905
M8	236	3.76	0	5	-0.701	-0.342
M9	236	3.94	0	5	-0.917	0.460
M10	236	4.12	0	5	-1.43	1.63
M11	236	3.62	0	5	-0.771	0.115
M12	236	4.01	0	5	-1.04	0.949
M13	236	3.75	0	5	-0.68	-0.421
M14	236	3.21	0	5	-0.419	-0.896
M15	236	3.87	0	5	-0.964	0.868
RTS	236	5.95	4.0	8.0	-0.305	4.32

The collected data was imported into AMOS 24 to check the model fit of the proposed model. After the initial analysis, the model fit of the proposed model was not acceptable, and then the appropriate modification was taken. For the model modification, first, the researchers allowed the covariances of items M.I. over 10 and within the same factor. After this modification, the model fit had changed, but was still not acceptable. Then, the researchers decided to delete items which standardised residual covariance was over two and removed these observed variables. A review of the observed variables confirmed that the removal of items M1, M2, M6, and M11 was justified both statistically and theoretically. Theoretically, M1 and M2 focused on *passive* self-assessments of inability (e.g., "English is not good enough"), which overlapped heavily with the latent construct of anxiety in V3 (Individual Differences). By removing these, the factor V1 (Self-perceived Proficiency) was refined to focus more precisely on active learning needs. Similarly, removing M6 and M11 reduced redundancy within the Academic Expectations and Individual Differences constructs. This refinement ensured that the remaining observed variables provided a distinct and comprehensive representation of each latent construct, thereby upholding the construct validity of the proposed model as recommended by Kline (2016). After checking the description of each observed variables, the researchers found the removal of these factors and keep the rest of the factors did not threaten the underlined construct each latent variable and the overall

construct of the model would not be threatened. In addition, for the statistical analysis of SEM, there should be at least 3 observed variables of each latent variable as this could allow for a more comprehensive representation of the latent construct, which improves the construct validity of the proposed model (Kline, 2016). By considering all the evidence, the researchers found that the removal of these factors did not have an impact on the construct validity of the model.

After these two steps, a further analysis was conducted to check the model fit of the three factors model. MLR was employed to estimate model parameters and goodness-of-fit of the modified three-factor model with: RMSEA < 0.06 (90% CI ≤ 0.06), SRMR, CFI=≥ 0.95, and TLI =≥ 0.95 (Hu & Bentler, 1999; Brown, 2015). Additionally, the chi-square/df ratio ≤ 3 rule was also used (Kline, 2016). Based on the previous studies and PCA that were performed in the previous phase, the following model 1 (Figure 6) was examined. In terms of the model fit, model 1 showed an acceptable fit (Table 4).

Table 4: Model fit indexes

Model fit indicts	Suggested	Obtained	References
RMSEA	< 0.06(90% CI ≤ 0.06)	0.051	Hu & Bentler, (1999); Brown (2015)
SRMR	≤ 0.08	0.08	
CFI	≥ 0.95	0.966	
TLI	≥ 0.95	0.954	
Chi-square/df ratio	≤ 3	1.692	Kline (2016)

Regarding the factor loadings of each observed variable (e.g., M1, M2, M3) to the factors (e.g., V1, V2, V3), all of them were over 0.4, and this could be interpreted as all the variables could at least moderately explain the associated factor (Hair et al., 1998). The model fit suggests construct validity, which test-takers' reasons for undertaking a test preparation course and the IELTS speaking score has a connection. In addition to the model fit, the regression weights (estimates), standard errors (S.E.), and p-values for the effects of V1, V2, and V3 on the test score (RTS) should also be shown (Table 5).

The non-significant results suggested that the specific reasons (V1, V2, and V3) for test-takers' undertaking the test preparation course did not significantly and directly affect their test scores. This finding is consistent with previous studies, which indicate that washback on learning is a complex phenomenon influenced by multiple factors, either directly from the test or by construct-irrelevant variances (Green, 2007; Xie, 2013; Dong, 2020; Zhang & Bournot-Trites, 2021). However, the lack of significant effects from these individual factors may suggest that other factors or a combination of factors might be more important in explaining the mechanism of the washback effect on learning. In other words, the quantitative findings from the current study could imply that the connection between test-takers' reasons for undertaking a test preparation course and test scores might be indirect. This finding offers empirical evidence to argue that mediating factors should no longer be ignored in washback on learning studies or considered solely as construct irrelevant variance. Rather, taking into account these mediating factors can significantly enhance our understanding of how the interplay between these factors influences learning behaviour and, potentially, test scores through indirect pathways.

Figure 6: SEM output Model

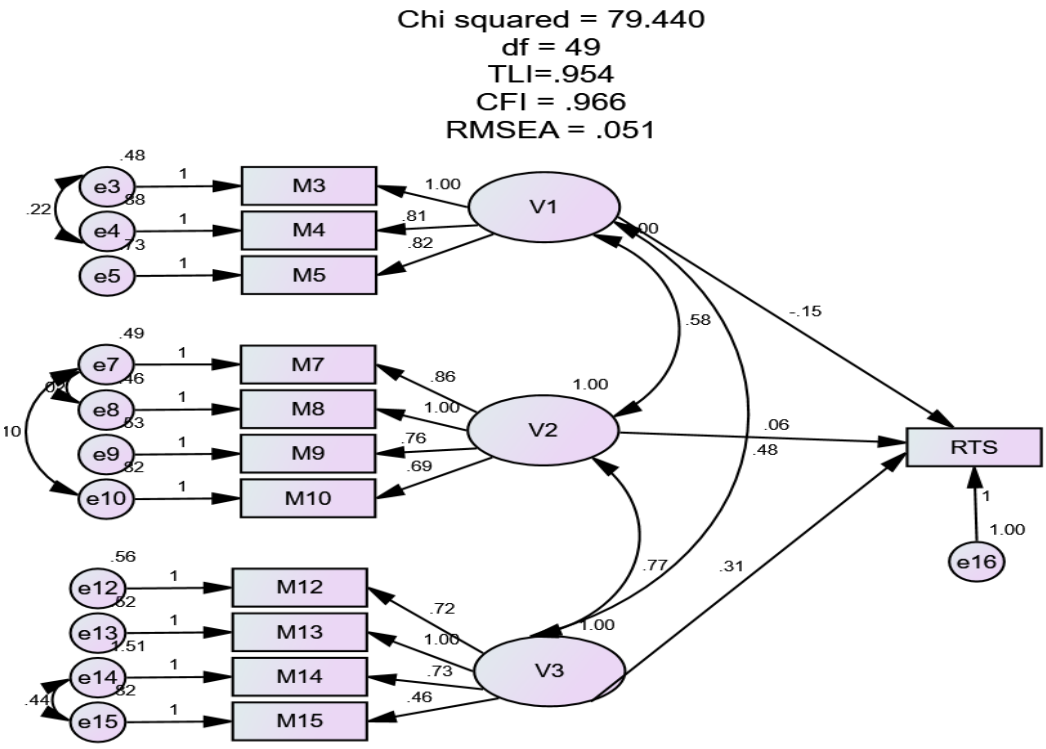


Table 5: Descriptive statistics of S.E. and p value

Latent Variable	Estimate	S.E.	C.R.	P-value	Significance
V1	-0.248	0.138	-1.792	0.073	Not significant (p > 0.05)
V2	0.367	0.197	1.863	0.062	Not significant (p > 0.05)
V3	0.121	0.168	0.723	0.470	Not significant (p > 0.05)

4.3. Qualitative Analysis

Following each closed-ended questionnaire item, participants were invited to provide open-ended responses. This allowed them to share personal reflections and explanations, offering richer insights into their perspectives on the IELTS Speaking Test and their reasons for enrolling in preparation courses. An inductive approach was applied to analyse these responses.

Recurring words, phrases, or ideas mentioned at least twice were coded and organised into themes. Two tables summarise the findings: Table 6 presents participants' interpretations of the IELTS Speaking Test design, while Table 7 highlights their reasons for undertaking preparation courses.

Table 6: Interpretation of the test design

Factors	Open-ended questions	Responses
Interpretation of the test construct	What other abilities do you think the IELTS speaking test part 1 assesses?	The ability to express oneself ideas Thinking skills: the answers should be logical and concise The ability to react according to different situations or questions Listening skill
	What other abilities do you think the IELTS speaking test part 2 assesses?	Ability to give logical answers Ability to express one's own idea Ability to organize language to make a presentation Ability to make a fluent speech
	3．What other abilities do you think the IELTS speaking test part 3 assesses?	Critical thinking skill The ability to express your own ideas The ability to make summaries reasoning ability knowledge base
	4．What grammar and vocabulary knowledge do you think the whole speaking test assesses?	Use more complex grammar structures Can use correct grammar and appropriate grammar structures Use more words that are less-frequently used in the conversation
	What other ways would you choose to form your answers?	Prepare materials for each topic in advance, and select and paraphrase materials to answer questions. Recite materials that can answer various questions. Performance on spot
Self-perceived evaluation goals	What other criteria do you think the IELTS speaking test will use to assess your test performance?	Being resourceful Using correct grammar and daily language Vocabulary size Ability to express one's idea Flow of speech
	What other abilities do you believe the IELTS speaking test plan to evaluate in the real test?	Logical thinking Ability to organize language Ability to think in English (no Chinglish/ use authentic English words/ expression) Ability to communicate in English (listening and speaking)

Table 7: Reasons for undertaking a test preparation course

Factors	Open-ended questions	Responses
Self-perceived language proficiency	What do you think about your English-speaking ability level?	Moderate / average Being able to communicate in English

Academic expectations	What other kinds of expectation do you have from taking an IELTS preparation course?	pass the test/ get the score I want Improving my spoken English help me build up my learning habits and the school could provide more learning materials being able to express ideas confidently in English learning foreigners' logical thinking pattern
Individual differences	What other personal factors do you think will affect your attitudes to take an IELTS preparation course?	Limited time to prepare the test by myself / limited energy to handle many things at the same time Feeling not confident to take the test Feeling not prepared enough to take the test

4.3.1. Interpretation of the test design

Participants offered a wide range of views on the abilities assessed in the IELTS Speaking Test. Their comments were grouped into three thematic strands, coded as black, blue, and red.

Black strand refers to the test-takers' alignment with test design. To be more specific, many responses closely mirrored the skills explicitly targeted by IELTS, such as expressing ideas clearly, demonstrating fluency and coherence, and using accurate grammar and vocabulary. This alignment suggests that a number of participants held an appropriate understanding of the test construct and its assessment criteria.

The 'blue strand' represents additional skills perceived by test-takers. Some participants believed the test assessed broader abilities not explicitly stated in the official materials, including critical thinking, reasoning, knowledge base, logical thinking, and organisational skills. While these abilities are not directly measured, participants viewed them as integral to strong performance. This perception aligns with the IELTS framework as a communicative test (IELTS, 2021), where higher-order cognitive abilities naturally interact with communicative competence.

Red strand, this category refers to test-takers' misinterpretations of the skills the test aims to assess. A subset of participants demonstrated a misunderstanding about what the test rewards. For example, some believed complex grammar and rare vocabulary would lead to higher scores, despite IELTS guidelines emphasising *accuracy* and *appropriacy* (IELTS, 2020). Others described preparing by rote memorisation and recitation—strategies reflecting the prevalence of passive learning among Chinese learners (Sit, 2013; Huang & Cowden, 2009). While culturally familiar, these approaches may disadvantage test-takers in a communicative performance-based exam.

In summary, participants' interpretations ranged from accurate alignment to partial misalignment or expansion of the test construct. These findings illuminate why discrepancies emerged in the quantitative results: cultural learning traditions and preparation practices shaped beliefs about what the test actually measures.

4.3.2. Reasons for Taking a Preparation Course

Participants also explained their motivations for enrolling in IELTS preparation courses. These reasons clustered around three mediating factors:

i. Self-perceived language proficiency: Many participants described themselves as having *moderate* English ability and being able to communicate in everyday contexts. However, quantitative results indicated that they lacked confidence in their speaking skills and feared these would not be sufficient to achieve their target scores.
ii. Academic expectations: Participants expressed both test-oriented and learning-oriented goals. Some sought primarily to "pass the test" or "achieve a target score," while others emphasized improving spoken English, gaining study materials, developing learning habits, and building confidence in English expression. A few mentioned learning from *foreigners' logical thinking patterns*. These findings nuance earlier research, which has tended to portray preparation courses as primarily goal-oriented (Popham, 1997; Crocker, 2006).
iii. Individual differences: Beyond proficiency and goals, participants highlighted personal factors influencing their decision to enrol. These included limited time and energy to self-study, lack of confidence, and feelings of unpreparedness. For many, a preparation course was seen as a way to ease the burden of study and maximise efficiency.

Taken together, the qualitative data largely reinforced the quantitative results while also providing explanatory depth. Participants' beliefs about the test design often diverged from the intentions of test developers. Some accurately recognised assessed skills, while others misinterpreted or expanded the construct to include additional abilities such as reasoning and knowledge base. This misalignment underscores the importance of clearer communication from test developers and educators about the skills being assessed.

Furthermore, participants' reasons for undertaking preparation courses revealed that their motivations were not exclusively test-focused. While many sought to achieve specific scores, others valued skill development, learning strategies, and confidence-building. This suggests that preparation courses play both an instrumental and a developmental role for learners.

In conclusion, these qualitative insights emphasise the need for preparation materials and curricula that: 1. Clearly define the intended construct and assessment criteria of the IELTS Speaking Test; 2. Emphasise communicative competence over rote memorisation; 3. Support both goal-oriented and learning-oriented student needs.

5. Discussion

5.1. RQ1: What is the relationship between Chinese IELTS test-takers' interpretation of the IELTS speaking test design and the perceptions of the intended washback?

One of the most salient findings of this study is the significant misalignment between test-takers' interpretations of the IELTS Speaking Test design and the official assessment criteria. Quantitatively, participants demonstrated a limited understanding of the test

construct but reported a high level of self-perceived understanding. Qualitatively, this discrepancy manifested as a *perception–action gap*: many test-takers believed they understood the test's evaluation goals, yet their preparation strategies often contradicted the intended communicative nature of the assessment.

This misalignment offers an explanation for the paradox observed in the Chinese EFL context, high participation in preparation courses but limited improvement in speaking scores. Learners' efforts are substantial, but their understanding of what the test measures is often misguided. As a result, they engage in activities that do not contribute effectively to the development of the skills valued by the IELTS Speaking Test.

This finding resonates with Sato (2018) and Dong (2020), who found that learners' misinterpretations of test constructs can lead to counterproductive preparation behaviours. From a validity perspective, it also supports Kane's (2013) argument that the consequential aspect of validity must consider not only the intended effects of testing but also the unintended, learner-level consequences that arise from misunderstanding test purposes. In this sense, the observed misalignment represents a consequential validity concern: when learners' perceptions of the construct diverge from the test developers' intentions, the test may inadvertently promote ineffective or even negative washback.

To promote positive washback, test developers and educators must therefore address this interpretive gap. Initiatives such as visual rubrics, training videos, or orientation workshops could clarify assessment criteria, helping test-takers to align their preparation behaviours with the communicative objectives of the test.

5.2. RQ2: How could the interplay between the mediating factors have an impact on the intended washback of the IELTS speaking test on test-takers' learning behaviours and test score?

The second research question investigated the role of mediating factors—self-perceived proficiency, academic expectations, and individual differences—in shaping test-takers' preparation behaviours and test outcomes. The SEM results revealed no significant *direct* relationship between these mediating variables and test scores. While this may initially suggest that personal factors do not impact performance, a deeper theoretical interpretation is required. These non-significant direct paths validate the distinction in Bailey's (1996) model between *Processes* (preparation behaviours) and *Products* (scores). The findings suggest that mediating factors (V1–V3) exert their influence on the *process* of learning—dictating whether a student chooses rote memorisation or communicative practice—rather than directly causing the *product* (the score). For instance, the qualitative data showed that students with high anxiety (Individual Difference) often resorted to "reciting materials," a strategy that reduces immediate stress but fails to improve the Speaking score. Therefore, the washback effect here is functional but indirect: mediating factors shape the *quality* of preparation, which subsequently determines the score, rather than influencing the score directly.

This interpretation is consistent with Green's (2007) and Dong's (2020) findings that the influence of washback operates through mediated pathways involving motivation, perception, and self-regulation. The current study therefore extends Bailey's (1996) model by empirically substantiating the complexity of the *"Processes"* component. Whereas Bailey's (1996) original framework described processes as learning behaviours

arising from test influence, this study demonstrates that those processes are themselves *mediated* by learners' perceptions and personal factors.

In this way, the present research supports the reconceptualisation of washback as a non-linear, multi-level mechanism, consistent with Cheng et al. (2015) and Allen (2016). It also reinforces the argument that mediating factors—traditionally dismissed as construct-irrelevant variance should instead be viewed as integral to understanding test consequences (Messick, 1996; Bachman, 2005). Rather than compromising test validity, these factors reveal the contextual realities through which validity claims must operate.

6. Contributions and Implications

This study makes two theoretical contributions to the understanding of the mechanism of washback on learning. First, it provides empirical support for the mediated nature of Bailey's (1996) model, emphasising that the *Processes* component—learning and preparation behaviours—is shaped by dynamic interactions between test-taker perceptions and contextual mediators.

Second, it contributes to the consequential validity literature (Messick, 1989; Kane, 2013; Bachman, 2005; Chalhoub-Deville & O'Sullivan, 2020) by evidencing how learners' misinterpretation can generate unintended negative consequences. The test itself may be well-constructed, but if its intended construct is misunderstood by learners, the resulting preparation behaviours may undermine the very communicative competencies the test seeks to promote.

In addition to contributions, this study offers practical implications to different stakeholders. For IELTS test developers, proactive communication with stakeholders is essential. For example, the development of short, accessible video series for test-takers that explicitly debunk common myths (e.g., "You don't need complex words to get a high score") or explain the rationale behind the speaking test criteria can be helpful to pay more attention to the intended evaluation goals of the target test. For test preparation course designers, conducting needs analyses to tailor instruction. For students focused on scores, provide strategic feedback against the rubric. For those seeking skill development, design authentic, communicative tasks.

7. Limitations

This study is subject to several limitations that should be acknowledged when interpreting the findings. First, the use of convenience sampling limits the representativeness of the participant group. While the sample size of 236 provides adequate statistical power, it may not fully capture the diversity of IELTS test-takers across China in terms of region, proficiency level, or educational background. Future research could employ stratified random sampling to enhance generalisability. Second, the study relied primarily on self-reported questionnaire data, which may be influenced by recall bias or social desirability effects. Although the inclusion of open-ended questions provided rich qualitative insights, triangulating these with classroom observations or interview data could offer a more comprehensive understanding of test preparation behaviours. Third, the cross-sectional design limits the ability to infer causal relationships between mediating factors, preparation behaviours, and test scores. Longitudinal studies could better capture the evolving nature of washback as learners progress through different stages of test preparation. Fourth, this study focused

exclusively on the IELTS Speaking Test. Given that washback can differ across test components (e.g., reading, writing, listening), future research should explore whether similar perceptual and behavioural patterns emerge in other sub-tests. Finally, while the current study situates Bailey's (1996) model within the Chinese EFL context, further cross-cultural validation is warranted. Comparative research in East Asia or non-Anglophone contexts could reveal whether the mediating mechanisms identified here are culturally specific or more universally applicable.

8. Conclusion

This study demonstrates that the mechanism of the IELTS speaking test washback on Chinese learners is characterized by a critical misalignment between perception and practice. Despite high confidence in their understanding of the test, participants often misinterpreted its communicative goals, leading to preparation strategies centered on rote memorization rather than the development of genuine spoken skills. This perceptual gap directly explains the paradox of high course enrolment but limited score improvement.

Furthermore, the research establishes that the influence of key mediating factors, such as self-perceived proficiency and individual differences is indirect. These factors do not directly determine test scores but instead shape the quality of learning behaviours, often channelling learners toward ineffective preparation methods. Consequently, the study affirms that washback is a complex, process-oriented phenomenon, underscoring the urgent need for pedagogical interventions that correct misconceptions and promote the authentic communication skills the test is designed to measure.

Ethics Approval and Consent to Participate

The researchers used the research ethics provided by Universiti Sains Malaysia (USM) Human Research Ethics Committee (JEPeM). All procedures performed in this study involving human participants were conducted in accordance with the ethical standards of the institutional research committee. Informed consent was obtained from all participants according to the Declaration of Helsinki.

Acknowledgement

Part of this article was extracted from a doctoral thesis submitted to Universiti Sains, Malaysia, Penang.

Funding

This study received no funding.

Conflict of Interest

The authors reported no conflicts of interest for this work and declare that there is no potential conflict of interest with respect to the research, authorship, or publication of this article.

References

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115–129.

Allen, D. (2016). Investigating washback to the learner from the IELTS test in the Japanese tertiary context. *Language Testing in Asia, 6*(1), 1–20. https://doi.org/10.1186/s40468-016-0030-z

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1–34. https://doi.org/10.1080/15434300590941154

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

Booth, D. K. (2018). *The sociocultural activity of high stakes standardised language testing: TOEIC washback in a South Korean context* (Vol. 12). Springer.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Publications.

Chalhoub-Deville, M., & O'Sullivan, B. (2020). *Validity: Theoretical development and integrated arguments*. British Council.

Chappell, P., Yates, L., & Benson, P. (2019). *Investigating test preparation practices: Reducing risks* (IELTS Research Reports Online Series, No. 3). British Council, Cambridge Assessment English, and IDP: IELTS Australia.

Cheng, L. (2005). *Changing language teaching through language testing: A washback study*. Cambridge University Press.

Cheng, L., & Deluca, C. (2011). Voices from test-takers: Further evidence for language assessment validation and use. *Educational Assessment, 16*(2), 104–122. https://doi.org/10.1080/10627197.2011.584042

Cheng, L., Sun, Y., & Ma, J. (2015). Review of washback research literature within Kane's argument-based validation framework. *Language Teaching, 48*(4), 436–470. https://doi.org/10.1017/S0261444815000233

Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and conducting mixed methods research* (5th ed.). Sage Publications.

Crocker, L. (2006). Preparing Examinees for Test Taking: Guidelines for Test Developers and Test Users. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 115–128). Lawrence Erlbaum Associates Publisher

Dong, M. (2020). Structural relationship between learners' perceptions of a test, learning practices, and learning outcomes: A study on the washback mechanism of a high-stakes test. *Studies in Educational Evaluation, 64*, 100824. https://doi.org/10.1016/j.stueduc.2019.100824

Ferman, I. (2004). The washback effect of an EFL national oral matriculation test on teaching and learning. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 191–210). Lawrence Erlbaum Associates.

Fox, J., & Cheng, L. (2007). Did we take the same test? Differing accounts of the Ontario secondary school literacy test by first- and second-language test takers. *Assessment in Education: Principles, Policy & Practice, 14*(1), 9–26. https://doi.org/10.1080/09695940701272773

Gan, Z. (2009). IELTS preparation course and student IELTS performance: A case study in Hong Kong. *Journal of Language Teaching and Research, 40*(1), 23–41. https://doi.org/10.4304/jltr.40.1.23-41

Gosa, C. M. C. (2004). *Investigating washback: A case study using student diaries* [Doctoral dissertation, University of Lancaster].

Green, A. (2007). Washback to learning outcomes: A comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education: Principles, Policy & Practice, 14*(1), 75–97. https://doi.org/10.1080/09695940701272880

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (1998). *Multivariate data analysis* (5th ed.). Prentice Hall.

Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. *Language Testing, 14*(3), 295–303. https://doi.org/10.1177/026553229701400306

Hawkey, R. (2006). *Impact theory and practice* (Studies in Language Testing 24). Cambridge University Press.

Hayes, B., & Read, J. (2004). IELTS test preparation in New Zealand: Preparing students for the IELTS academic module. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing* (pp. 119–134). Routledge.

He, S., Sénécal, A. M., Stansfield, L., & Suvorov, R. (2024). A scoping review of research on second-language test preparation. *Language Testing, 41*(1), 1–28. https://doi.org/10.1177/02655322231190788

Horwitz, E. K. (2010). Foreign and second language anxiety. *Language Teaching, 43*(2), 154–167. https://doi.org/10.1017/S0261444809990376

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hu, R., & Trenkic, D. (2021). The effects of coaching and repeated test-taking on Chinese candidates' IELTS scores, their English proficiency, and subsequent academic achievement. *International Journal of Bilingual Education and Bilingualism, 24*(10), 1486–1501. https://doi.org/10.1080/13670050.2019.1691498

Huang, J., & Cowden, P. (2009). Are Chinese students really quiet, passive and surface learners? A cultural studies perspective. *Comparative and International Education, 38*(2), 1–15.

Hughes, A. (1993). *Backwash and TOEFL 2000* [Unpublished manuscript]. University of Reading.

IELTS. (2019). *IELTS performance for test-takers 2019*. https://www.ielts.org/teaching-and-research/test-taker-performance

IELTS. (2020). *IELTS test format*. https://www.ielts.org/about-the-test/test-format#tab-6

IELTS. (2021). *IELTS performance for test-takers 2021*. https://www.ielts.org/for-researchers/test-statistics/test-taker-performance

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/j.1745-3984.2012.00264.x

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Publications.

Larson-Hall, J. (2015). *A guide to doing statistics in second language research using SPSS and R*. Routledge.

Ma, H., & Chong, S. W. (2022). Predictability of IELTS in a high-stakes context: A mixed methods study of Chinese students' perspectives on test preparation. *Language Testing in Asia, 12*(1), 1–18. https://doi.org/10.1186/s40468-021-00152-3

Ma, J. (2017). *Understanding test preparation phenomenon through Chinese students' journey towards success on high-stakes English language tests* [Doctoral dissertation, Queen's University]. QSpace.

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. *Psychological Bulletin, 89*(3), 575–588. https://doi.org/10.1037/0033-2909.89.3.575

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11. https://doi.org/10.3102/0013189X018002005

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241–256. https://doi.org/10.1177/026553229601300302

Mickan, P., & Motteram, J. (2009). *The preparation practices of IELTS candidates: Case studies* (IELTS Research Reports, 10). IELTS Australia.

Mizutani, S. (2009). *The mechanism of washback on teaching and learning* [Unpublished doctoral thesis]. University of Auckland.

Popham, W. J. (1997). Consequential validity: Right concern—wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9–13. https://doi.org/10.1111/j.1745-3992.1997.tb00442.x

Sato, T. (2018). The impact of the Test of English for Academic Purposes (TEAP) on Japanese students' English learning. *JACET Journal, 62*, 89–107.

Sato, T. (2019). An investigation of factors involved in Japanese students' English learning behaviour during test preparation. *Papers in Language Testing and Assessment, 8*(1), 69–95. https://doi.org/10.1075/plta.8.1.04sat

Shih, C. (2007). A new washback model of students' learning. *Canadian Modern Language Review, 64*(1), 135–161. https://doi.org/10.3138/cmlr.64.1.135

Sit, H. H. W. (2013). Characteristics of Chinese students' learning styles. *International Proceedings of Economics Development and Research, 62*, 36–40.

Tsang, C. L. H. (2017). *Examining washback on learning from a sociocultural perspective: The case of a graded approach to English language testing in Hong Kong* [Unpublished master's thesis]. University College London.

Tsang, C. L., & Isaacs, T. (2022). Hong Kong secondary students' perspectives on selecting test difficulty level and learner washback: Effects of a graded approach to assessment. *Language Testing, 39*(2), 212–238. https://doi.org/10.1177/02655322211050600

Wall, D. (1997). Impact and washback in language testing, *Encyclopedia of language and education,* 7, 291-302.

Wei, W. (2017). A critical review of washback studies: Hypothesis and evidence. In *Revisiting EFL assessment* (pp. 49–67). Springer.

Xie, Q. (2011). Is test taker perception of assessment related to construct validity? *International Journal of Testing, 11*(4), 324–348. https://doi.org/10.1080/15305058.2011.589018

Xie, Q. (2013). Does test preparation work? Implications for score validity. *Language Assessment Quarterly, 10*(2), 196–218. https://doi.org/10.1080/15434303.2012.721423

Xie, Q., & Andrews, S. (2013). Do test design and uses influence test preparation? Testing a model of washback with structural equation modeling. *Language Testing, 30*(1), 49–70. https://doi.org/10.1177/0265532212442634

Yu, G., He, L., Rea-Dickins, P., Kiely, R., Lu, Y., Zhang, J., Zhang, Y., Xu, S., & Fang, L. (2017). *Preparing for the speaking tasks of the TOEFL iBT® test: An investigation of the journeys of Chinese test takers* (ETS Research Report Series). Wiley.

Zhan, Y., & Wan, Z. H. (2016). Test takers' beliefs and experiences of a high-stakes computer-based English listening and speaking test. *RELC Journal, 47*(3), 363–376. https://doi.org/10.1177/0033688215626498

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. https://doi.org/10.1016/j.stueduc.2021.100977