



Castledown

 OPEN ACCESS

Language Education & Assessment

ISSN 2209-3591

<https://www.castledown.com/journals/lea/>

Language Education & Assessment, 9, 103375 (2026)

<https://doi.org/10.29140/lea.2026.103375>

Examining the Mechanism of Washback on Learning of the IELTS Speaking on Chinese EFL Context



YUAN LIU^a 

ALLA KHAN^b 

^a*Universiti Sains Malaysia, China*

yuan.liu7@outlook.com

^b*Universiti Sains Malaysia, Malaysia*

allabaksh@usm.my

Abstract

This study explores the mechanism of washback of the IELTS speaking test on learning in the Chinese EFL context by examining the interplay between IELTS speaking test-takers' interpretation of test design and their reasons for enrolling in test preparation courses, and how these factors relate to speaking scores. Using Bailey's (1996) Washback Model and a mixed-methods design, data were collected from 267 Chinese IELTS test-takers through surveys and 15 semi-structured interviews. Quantitative findings from confirmatory factor analysis (CFA) and structural equation modelling (SEM) revealed that both test design interpretation and mediating factors (self-perceived language proficiency, academic expectations, and individual differences) were associated with speaking scores. Qualitative analysis highlighted misalignments between perceived and intended test constructs. The study proposes a contextualised model of washback that operationalises the 'process' component as the dynamic interplay between test-takers' interpretation and mediating factors, contributing to our understanding of learner behaviour and offering practical implications for test preparation pedagogy.

Keywords: Washback on learning, IELTS speaking test, test preparation course, test scores, Chinese EFL context

Introduction

In recent years, the growing demand for international education among Chinese students has significantly increased the importance of high-stakes English proficiency tests, most notably the International

Copyright: © 2026 Yuan Liu & Alla Khan. This is an open access article distributed under the terms of the Creative Commons Attribution Non-Commercial 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within this paper.

English Language Testing System (IELTS). Among its four components, the IELTS Speaking test remains particularly challenging for Chinese English as a Foreign Language (EFL) learners. Despite extensive participation in IELTS preparation courses, Chinese test-takers continue to underperform in the speaking component, with a national average score of just 5.5 (IELTS Partners, n.d.). This persistent underachievement raises critical questions about the effectiveness of preparation efforts and the underlying factors influencing test scores.

While early washback research foregrounded instructional practices and curriculum change, the past decade has seen a substantial shift toward learner-centred investigations, demonstrating that test-takers actively interpret assessments and mediate washback through their beliefs, motivations, and learning behaviours (Xie & Andrews, 2013; Zhan & Andrews, 2014; Xie, 2015; Dong, 2020; Zhang & Bournot-Trites, 2021; Zhang, 2024; Dong & Fan, 2025). This interpretive process influences how they select learning strategies, engage in preparation, and ultimately their test performance. Incorporating test-takers' voices can thus reveal important insights into the alignment or misalignment between test design, preparation behaviours, and test scores. This issue is particularly pronounced in the speaking component of the IELTS, which requires real-time, interactive language use. This is because the speaking test is inherently more complex than receptive skills assessments, as it demands higher-order cognitive and communicative abilities. As a result, learners' interpretations of the test construct and scoring criteria can significantly shape their preparation approaches (Green, 2007; Dong, 2020; Ma & Chong, 2022). Incorrect assumptions about the subject of the assessment can result in inadequate preparation, despite the fact that learners dedicate significant time and energy.

In addition to test-related factors, broader mediating factors, such as learners' self-perceived language proficiency, academic expectations, and individual differences also shape how learners prepare for high-stakes assessments (Fox & Cheng, 2007; Sato, 2019; Dong et al., 2023). While previous research has examined these factors in isolation, the interplay between test-takers' interpretation of test design and their personal reasons for enrolling in preparation courses and the connection with test scores has received limited empirical attention. This interplay may offer a more nuanced understanding of the mechanism through which washback influences learning, where personal factors filter the test interpretation, which then shapes preparation behaviours and connects with test scores.

To address this gap, the present study draws on Bailey's (1996) Washback Model and adopts a mixed-methods design to explore how Chinese IELTS test-takers interpret the speaking test, what mediating factors influence their decisions to undertake preparation courses, and how these factors, taken together, relate to speaking scores. The core hypothesis is that the washback effect of the IELTS speaking test is mediated by two interacting components, (1) Test-takers' interpretation of test design and contextual mediating factors, which collectively form the Process component that relates to the Products (IELTS speaking test scores).

The study is guided by the following research questions:

1. How do Chinese IELTS test-takers interpret the test design of the IELTS Speaking test?
2. What mediating factors influence their decision to undertake a test preparation course?
3. How does the interplay between test-takers' interpretation of the test design and their reasons for undertaking a preparation course relate to speaking scores?

By addressing these questions, this study aims to advance the theoretical and empirical understanding of washback on learning, especially in IELTS speaking test preparation course context. The findings

have important implications for test designers, educators, and policymakers seeking to ensure that test-takers' preparation efforts are meaningfully aligned with test constructs and intended learning outcomes.

Literature Review

The Washback Effect in Language Testing

The term “washback” in language testing refers to the influence of a test on teaching and learning. Alderson and Wall (1993, p.117) defined it as “the influence of testing on teaching and learning,” encompassing positive/negative and intended/unintended impacts. Pearson (1988) emphasised its backward-acting nature, as tests typically follow teaching/learning periods. Spolsky (1994) argued washback requires foreseeable impacts, but later research expanded it to include all test-related effects (Wall, 1997; Hamp-Lyons, 1997), framing it as part of a broader “impact” on educational systems.

In this study, washback on learning from test-takers' perspectives was explored. Wall (2000) noted test-takers mediate assessment-learning relationships by choosing learning activities based on their own perceptions. Collecting their views helps understand washback mechanisms, particularly how perceived evaluation goals align with test designers' intentions.

Washback in Language Testing: Definitions and Scope

The concept of *washback*, which is the influence of testing on teaching and learning, has been a long-standing focus in language assessment research (Alderson & Wall, 1993). Initially conceptualised as a bidirectional influence on classroom practices, washback has since evolved to include both intended and unintended effects, positive and negative outcomes, and broader societal impacts (Cheng & Curtis, 2004; Wall, 1997). Spolsky (1994) argued that washback effects are only meaningful if they are foreseeable, while later perspectives, such as Hamp-Lyons (2000), emphasised the ethical implications of such influence.

Recently, Chalhoub-Deville and O'Sullivan (2020) have urged researchers to move beyond psychometric concerns and explore the consequential validity of assessments by including learners' perspectives and experiences. This change has resulted in increased focus on the mechanism of washback, viewed not simply as a linear causal relationship but as a multifaceted process influenced by learners' interpretations, social factors, and cognitive perspectives.

Learner Interpretations of Test Design

Test-takers are not passive recipients of assessment policies. Rather, they actively interpret what a test requires and adjust their behaviour accordingly (Wall, 2000). This perspective has informed a growing body of research exploring how learners perceive the test construct, test criteria, and the intended evaluation goals of high-stakes tests.

For example, Cheng (1998) found Hong Kong students engaged in activities aligning with their perceived test goals. Hawkey (2006) observed IELTS learners prioritised test-oriented exercises based on evaluation goal perceptions. Mizutani (2009) found that Japanese learners' beliefs about a national exam shaped their preparation strategies and engagement levels. In recent studies, Sato (2019) linked Japanese students' exam preparation behaviours to their perceived evaluation goals. Chappell et al. (2019) revealed that unclear expectations in the speaking component led to learner frustration, rote memorisation, and negative attitudes—undermining both learning quality and test performance.

These findings underscore the importance of understanding how learners interpret test design, including whether they understand the target test construct, test criteria as well as the intended evaluation goals. Misaligned perceptions may produce ineffective preparation behaviours, jeopardising the test's validity and intended outcomes (Green, 2013; Heidari Vinchek et al., 2024).

Mediating Factors in the Washback Process

While the interpretation of the test design plays a central role, numerous mediating factors influence how learners experience washback. These include self-perceived language proficiency, academic goals, and individual differences, such as anxiety or confidence (Cheng et al., 2015; Dong, 2020; Zhang, 2021). The selection of key mediating factors in this study, namely, self-perceived language proficiency, academic expectations, and individual differences is grounded in a synthesis of theoretical models and empirical evidence.

Theoretically, these constructs are integral to prominent washback frameworks. For example, they are operationalised as the “intrinsic factors” and “extrinsic influences” outlined in Shih's (2007) model and give concrete definition to the “participant” characteristics and “processes” within Bailey's (1996) washback model. Empirically, their significant role in shaping learner behaviour is well-documented. Studies consistently show that self-perceived proficiency gaps drive enrolment in test preparation (Fox & Cheng, 2007; Gan, 2009; Zhou & Yoshitomi, 2019; Rahim et al., 2024). Self-perceived proficiency has been found to influence learners' engagement with test preparation. Fox and Cheng (2007) reported that ESL students with perceived language deficits were more likely to seek coaching. Gan (2009) observed that lower-achieving students in Hong Kong disproportionately relied on commercial test-prep courses. These findings are echoed in more recent work by Zhou and Yoshitomi (2019) and Rahim et al. (2024), who linked test-takers' perceived skill gaps to motivation and strategy use.

In addition, academic expectations mediate the intensity and focus of learning efforts (Xie & Andrews, 2013; Wang & Huang, 2020). This mediating factor can be seen as admission targets or scholarship requirements that can shape preparation behaviours. Xie and Andrews (2013) demonstrated that learners internalise institutional goals, which in turn mediate washback effects. Similarly, Wang and Huang (2020) found that TOEFL learners adjusted their preparation intensity based on perceived academic stakes. Individual differences can critically influence preparation strategies (Horwitz, 2010; Shih, 2007). This includes test anxiety, confidence, and learning styles, add another layer of complexity. Horwitz (2010) highlighted the cognitive toll of test anxiety, while Ferman (2004) and Shih (2007) pointed to the influence of social factors such as peer and parental expectations. Although prior research has often analysed these factors independently, their simultaneous inclusion in this study enables an investigation into their dynamic interactions. By doing so, this can address a critical gap in comprehending the complex washback mechanism on learning through which washback influences learning behaviours and the connection with test scores.

Gaps in the Literature

Despite the growing interest in washback on learning, existing research often examines test related features and mediating factors in isolation. While a few studies have begun to explore their interaction, there is a need for a more integrated understanding of how this interplay relates to actual test scores. Although Dong (2020) empirically modelled the pathway from learner perceptions to behaviours and learning outcomes, and Tsang and Isaacs (2022) examined learner perceptions within graded assessment systems, neither study explicitly operationalised the *process* component of washback as an interaction between test interpretation and multiple mediating factors within the IELTS speaking

preparation context. Conversely, studies focused on score gains (Gan, 2009; Zhan & Wan, 2016) rarely examined learners' interpretive processes.

This lack of integration limits our understanding of the mechanism of washback on learning, particularly how multiple factors converge to shape learning behaviours and influence scores. To address this gap, the current study proposes a contextualised model of washback based on Bailey's (1996) framework. This study moves beyond examining factors in isolation by focusing on the dynamic relationship between test-takers' interpretations, their reasons for enrolling in preparation courses (mediating factors), and their speaking test scores.

Theoretical Framework of this Study

Washback as a concept has been theorised through various models that seek to explain how tests influence teaching and learning (Bailey, 1996; Green, 2007; Green, 2013; Shih, 2007). Each framework emphasises different components of this relationship, ranging from societal factors to classroom practices and learner cognition. A clear theoretical grounding is essential for examining washback as a mechanism, rather than treating it as a loose collection of test-related influences (Cheng et al., 2015).

Shih's (2007) model (Figure 1), for example, categorises influences into extrinsic factors (e.g., society, school, family), intrinsic factors (e.g., individual differences, perceptions), and test factors (e.g., content, structure). While valuable in highlighting multiple domains, it has been critiqued for treating these categories as relatively discrete and for providing a limited explanation of how they interact to shape learner behaviour.

Additional frameworks, such as Green's (2007, 2013) washback-to-learning outcomes perspective, highlight the connection between preparation practices and learning outcomes. This study usefully highlights the outcomes of preparation, but does not fully explain how learners' interpretations of test design and contextual influences combine to drive those practices. Similarly, more recent accounts (e.g., Dong, 2020; Tsang & Isaacs, 2022) highlight the importance of learner perceptions, but they too stop short of presenting a coherent mechanism, failing to show how the interplay between different factors and the connection with test scores.

Bailey (1996) Basic washback model remains one of the most influential and comprehensive accounts. It distinguishes between participants (teachers, learners, test designers), processes (teaching and learning behaviours), and products (learning outcomes, scores). Crucially, Bailey emphasises the bidirectional relationships among these elements, allowing researchers to examine not only how tests influence learning but also how learner perceptions feedback into preparation behaviours and outcomes. Although the model has been criticised for its broad treatment of "processes" (Hamp-Lyons, 1997; Wall, 1997), this flexibility provides an opportunity for empirical studies to specify and test concrete processes relevant to a given context.

Additionally, the present study builds on Bailey's (1996) washback model by specifically operationalising the process component as the dynamic interaction between test-takers' interpretations of test design and mediating contextual variables, including self-perceived proficiency, academic expectations, and individual differences. This research integrates interpretive and motivational dimensions within a single mechanism, which differs from Shih (2007), who concentrated on separate intrinsic/extrinsic influences, and Green (2007, 2013), who employed an outcome-based model. The theoretical refinement presents washback as an interactive, cyclical system influencing learners' interpretation of test design and preparation behaviour, rather than a unidirectional cause-and-effect relationship. This operationalised framework is visually represented in Figure 3.



Figure 1 Washback model of students' learning (Shih, 2007, p. 151).

Methodology

Research Design

This study employed a mixed-methods triangulation design, integrating quantitative and qualitative data to examine the mechanism of washback on learning. The design allowed for simultaneous collection and analysis of data, enabling triangulation of findings to strengthen interpretive validity (Creswell & Creswell, 2017).

Sampling

The current study used a convenience sampling method with purposeful selection criteria to recruit participants for both the questionnaire and semi-structured interview. While convenience sampling

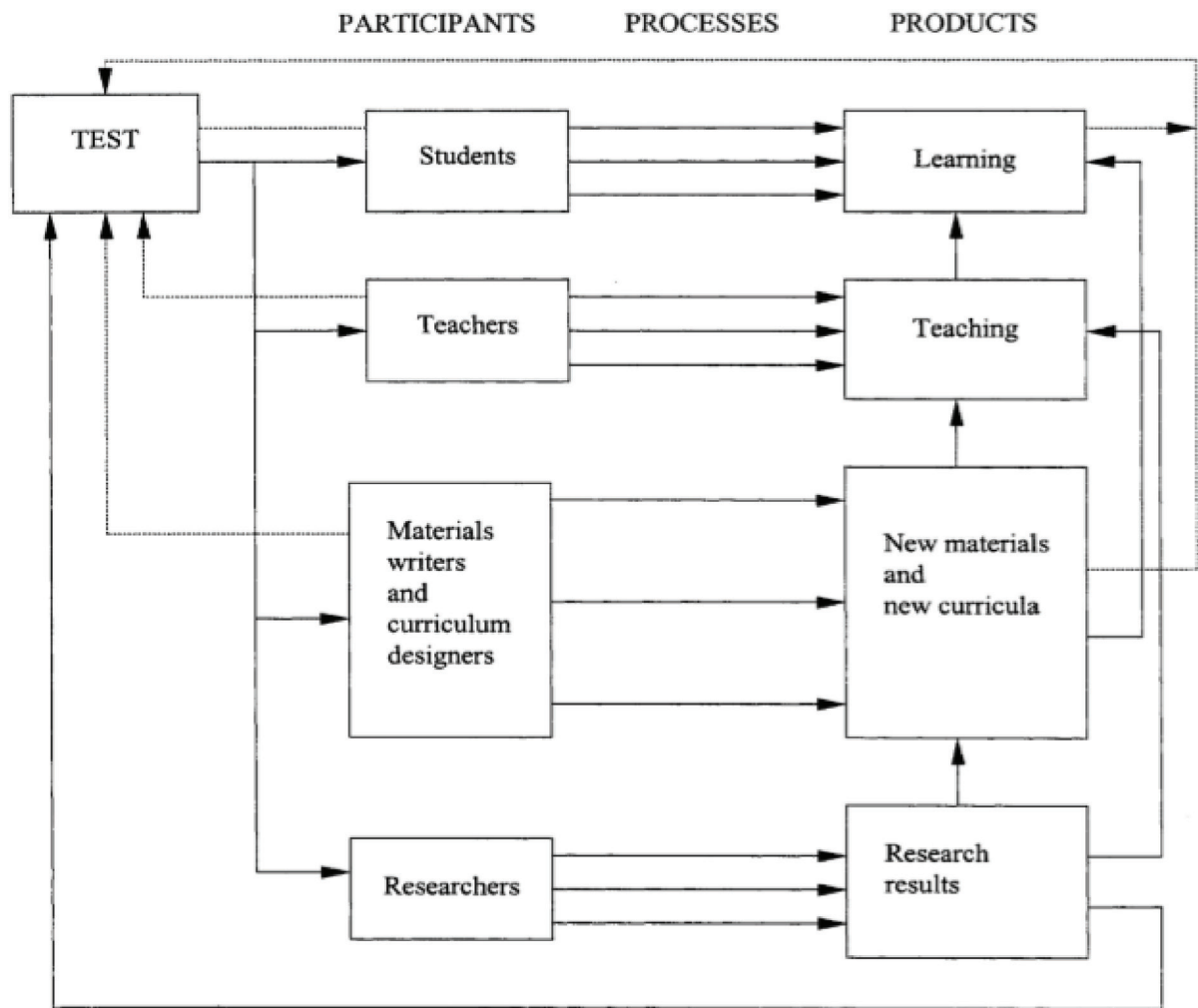


Figure 2 Basic model of washback (Bailey, 1996, p. 264).

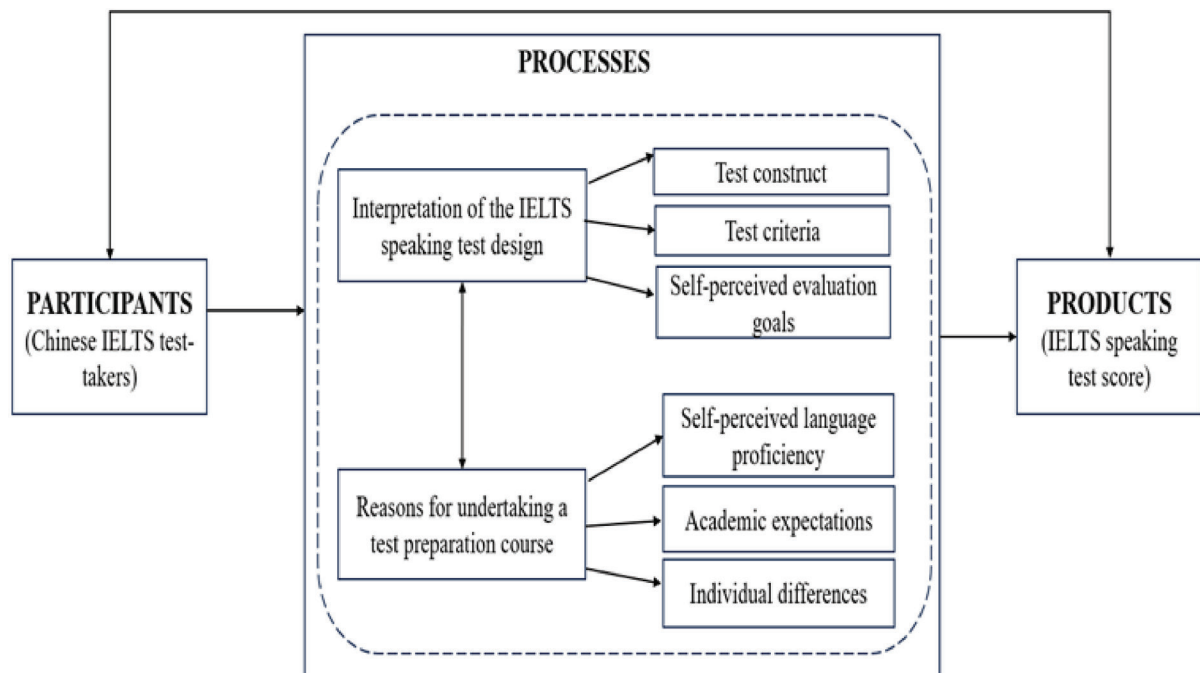


Figure 3 Conceptual model of washback on learning mechanism (adapted from Bailey, 1996).

permits a more focused investigation of a research topic, it also necessitates a clear understanding of the study's objectives and the characteristics of the relevant participants. The researcher chose participants for this study based on specific criteria, such as those who took an IELTS test preparation course and got their scores. The aim was to include individuals within the sample who could offer detailed, insightful perspectives on the phenomenon being researched. China Daily (2025) reports that mainland China, as the largest global market for IELTS, saw a record 1.04 million test-takers in 2023-2024. This extensive participant base facilitates the recruitment of volunteers for this study.

Participants

Quantitative phase: 267 Chinese IELTS test-takers were recruited through online platforms and university programs. Participants ranged from early 20s to early 30s, with a majority female sample, and were primarily university students or recent graduates. All had prior IELTS experience, had completed at least one commercial speaking preparation course, and reported IELTS speaking test band scores.

Qualitative phase: 15 participants from the survey sample were invited to join in the semi-structured interview (4 males and 11 females). For the qualitative analysis, participants were categorised into three groups based on their speaking scores they reported in the interviews to facilitate a comparative analysis of perspectives across proficiency levels: Group 1 (≥ 7.0 , $n = 3$), Group 2 (6.0–6.5, $n = 9$), and Group 3 (5.0–5.5, $n = 3$). This grouping aligns with common benchmarks in IELTS score interpretation and allows for examining how washback mechanism may differ between higher, intermediate, and lower-scoring test-takers.

Instruments

Quantitative data were collected via a structured questionnaire adapted from previous studies (Yu et al., 2017; Tsang, 2017; Iwashita et al., 2021; IELTS Partners, 2025), covering demographics, test design interpretation, and preparation course enrolment reasons. The validity of the questionnaire was confirmed via asking face validity from 5 test-takers and content validity with 2 TESOL experts as well as construct validity by conducting Principal Component Analysis (PCA), with Bartlett's test of sphericity ($p < 0.05$) and a KMO measure of 0.864, showing suitability for factor analysis and reliability (Cronbach's $\alpha = 0.858$). The final instrument comprised three primary sections. The first section collected participants' background information and their most recent IELTS Speaking scores. The second section was designed to elicit participants' interpretations of the test design, specifically their understanding of what the test aims to assess. This section employed multiple-choice questions (MCQs) to examine test-takers' interpretations of the test construct. For example, one item asked participants to identify the skills that Part 1 of the IELTS Speaking test is intended to assess. To capture self-perceived evaluation goals, a 5-point Likert scale (1 = Strongly disagree to 5 = Strongly agree) was used to measure the skills participants perceived the test to target. One sample item was, "I understand what specific speaking skills are being assessed in each part of the test." The third section was designed to examine test-takers' reasons for undertaking test preparation, which were categorised into self-perceived proficiency, academic expectations, and individual differences. These constructs were measured using a 5-point Likert scale.

Qualitative data were collected through semi-structured interviews that explored participants' interpretations of the IELTS Speaking test design and their reasons for undertaking test preparation courses. The interview protocol was developed to capture the nuances of the washback mechanism through open-ended questions. The protocol covered themes including test-takers' detailed views on the skills that Parts 1, 2, and 3 of the IELTS Speaking test aim to assess, their understanding of the test's scoring criteria, and their reasons for enrolling in test preparation courses. The interviews were guided by a

deductive thematic framework aligned with the study's conceptual model, and the data were analysed following Braun and Clarke's (2006) thematic analysis approach.

Data Collection

For the questionnaire, invitations were first distributed through contacts at IELTS preparation institutions. Second, the first author asked former students to share the study information with peers who were preparing for or had taken the IELTS test. Students who provided consent received the questionnaire link from the first author. Recruitment for the semi-structured interviews also employed two strategies: the first author contacted interested respondents from the questionnaire, and all authors posted invitations on social media platforms (WeChat and WhatsApp) to invite the participants. The interviews were then conducted by the first author using a structured interview protocol.

Ethical approval was obtained from the host university before data collection began. Prior to participation, all potential respondents were required to review an information letter and consent form. Participants provided informed consent, which included the right to withdraw at any time. All data were anonymised and stored securely (see appendix).

Data Analysis

For quantitative data, descriptive statistics (SPSS 24) analysed test design interpretation to answer research question 1. Amos 29 was employed to conduct confirmatory factor analysis (CFA) and structural equation modelling (SEM) to answer research question 2 and 3, respectively. In response to research question 3, which focuses on investigating the relationships between test-takers' interpretations of test design, mediating factors, and test scores, Structural Equation Modelling (SEM) was selected as the analytical approach. The measurement model (assessed via confirmatory factor analysis, CFA) was conducted first to validate the latent construct (RQ2). The full SEM was then used to test the hypothesised structural paths (RQ3). This approach allowed us to test a predefined relationship between the latent variables (representing the 'Processes' in Bailey's model) and the observed outcome variable (RTS, the 'Product'), directly addressing the concept of 'interplay' within the mechanism.

Initially, confirmatory factor analysis (CFA) was performed to validate the measurement model for the purpose of testing the hypothesised model, followed by full structural equation modelling (SEM). In the measurement model, the latent construct 'Interpretation of Test Design' (V1) was specified with three indicators: interpretation of test construct (I2), test criteria (I1), and self-perceived evaluation goals (MCQ). While these indicators represent distinct facets of knowledge, they were modelled as reflective indicators in the SEM analysis (Figure 6). This specification assumes that the underlying latent construct, the test-taker's holistic interpretation manifests through their understanding of these specific elements. Consequently, the arrows in the model point from the latent variable to the observed indicators, consistent with the reflective specification of the mediating factors (V2, V3, and V4).

For qualitative data, thematic analysis (NVivo 11) followed Braun and Clarke's (2006) framework. A deductive coding approach was employed, focusing on pre-determined themes aligned with the study's conceptual framework. Themes were refined iteratively to capture patterns related to test-takers' interpretations and preparation behaviours. This study used a single-coder approach for analysing qualitative data, which ensured consistency in coding and deeper familiarity with the dataset, leading to more nuanced insights (Freelon, 2010). However, this approach carries the risk of personal biases affecting coding and interpretation (O'Connor & Joffe, 2020). To address this, the researchers employed member checking, sharing preliminary findings with three participants to ensure accurate representation of their views. Additionally, NVivo 11 software was used to maintain coding consistency and reduce

human error (O’Kane et al., 2021). The study also used methodological triangulation, cross-checking qualitative findings with quantitative data and literature to enhance credibility (Tompkins et al., 2021). These strategies helped mitigate bias and improve the reliability of the findings.

Results

Quantitative Results

RQ1: How do Chinese IELTS test-takers interpret the test design of the IELTS speaking test?

The researchers employed the SPSS 24 software in order to gather and analyse the descriptive statistics regarding the test-takers’ interpretation of the speaking test design in Table 1, including their interpretation of the test construct, test criteria, and self-perceived evaluation goals. In terms of interpretation of the test construct, test criteria, all items were measured on a 1–5 Likert scale, where 5=Strongly Agree. There were five items (5 items/25) and six items (6 items/30) that constituted the test construct and test criteria, respectively.

The interpretation of test construct: Mean score = 20.9/25 (SD = 2.88), Test criteria interpretation: Mean score = 24.7/30 (SD=3.47), the relatively high standard deviation suggests divergent views on assessed skills. In terms of self-perceived evaluation goals, 5 multiple-choice questions (5 items / 5) were used to achieve this goal. The descriptive statistics of this section, Mean = 1.99/5 (SD = 0.99). A low standard deviation, alongside the low mean score, reveals a consistent perception among participants concerning the evaluation goals of the target test to be unfamiliar, which test-takers may hold their own interpretation of the evaluation goals that could be different from test developers aims to deliver to test users.

Table 1. *Descriptive statistics of the interpretation of test design*

	N	Mean	S. D
Test construct	267	20.9	2.88
Test criteria	267	24.7	3.47
Self-perceived evaluation goals	267	1.99	0.989

RQ2: What mediating factors influence their decision to undertake a test preparation course?

This question investigates how mediating factors, such as self-perceived language proficiency, academic expectations, and individual differences—influence test-takers’ learning behaviours, particularly their motivations for enrolling in test preparation courses. All the factors and observed variables listed in the Appendix and confirmatory factor analysis (CFA) model was developed (Figure 3) to visualise the relationships among these factors.

Prior to CFA, data normality was assessed via skewness (± 2 threshold) and kurtosis (± 7 threshold) tests, confirming normal distribution across all factors (Curran et al., 1996). Initial model fit was unsatisfactory, prompting refinements: allowing error term covariances (modification indices >10) within the same factor and deleting items with standardised residual covariances >2 . These adjustments preserved construct validity.

The modified three-factor model was estimated using maximum likelihood (ML) estimation. Model fit indices (Table 2) met or exceeded the recommended thresholds, indicating a good fit. Factor loadings

for all observed variables (e.g., M1–M3) exceeded 0.4, surpassing the moderate explanatory capacity threshold (Hair et al., 2010). The final CFA model (Figure 4) confirms that the three mediating factors significantly influence test-takers' learning behaviours, particularly in relation to preparation course enrolment.

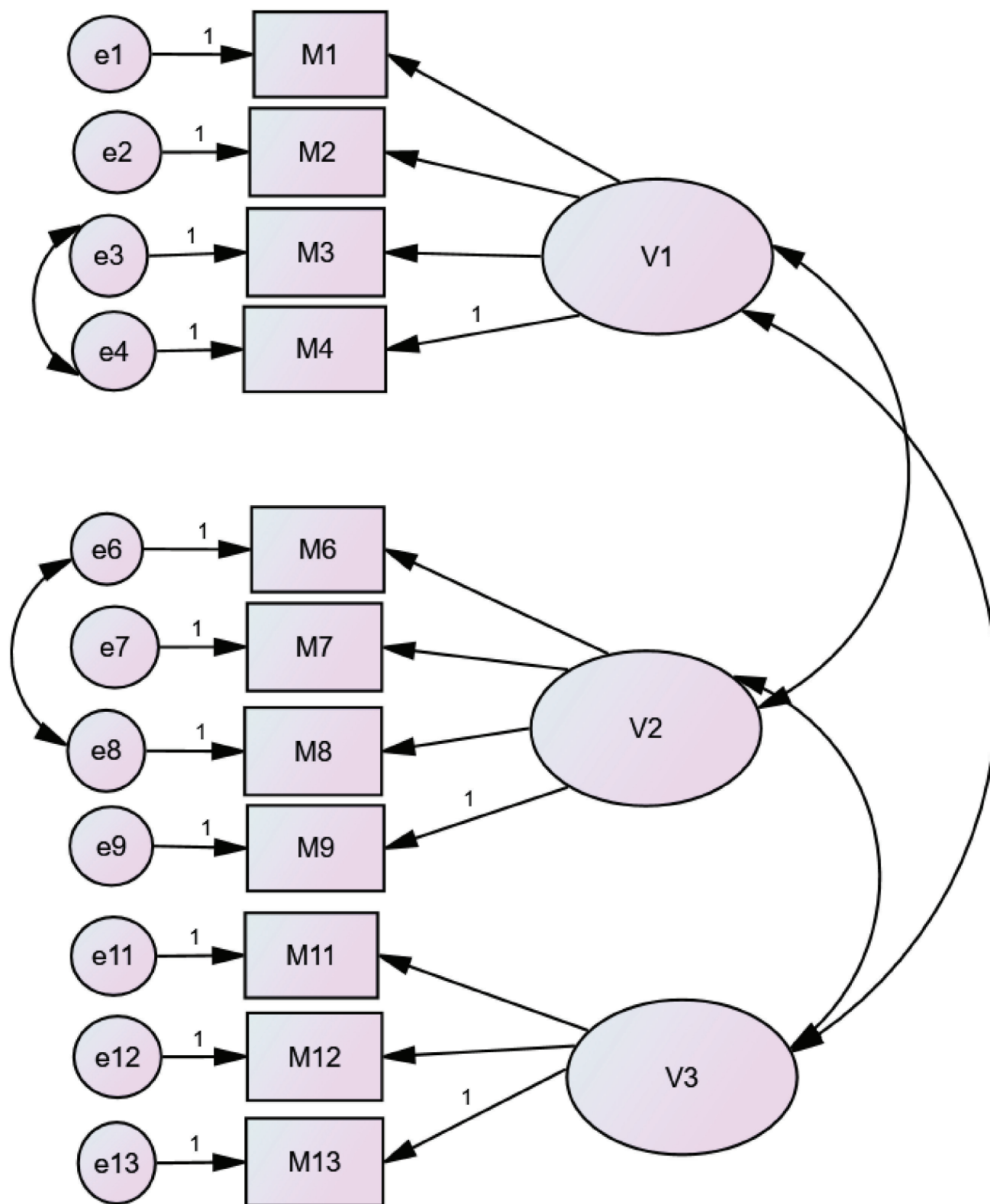
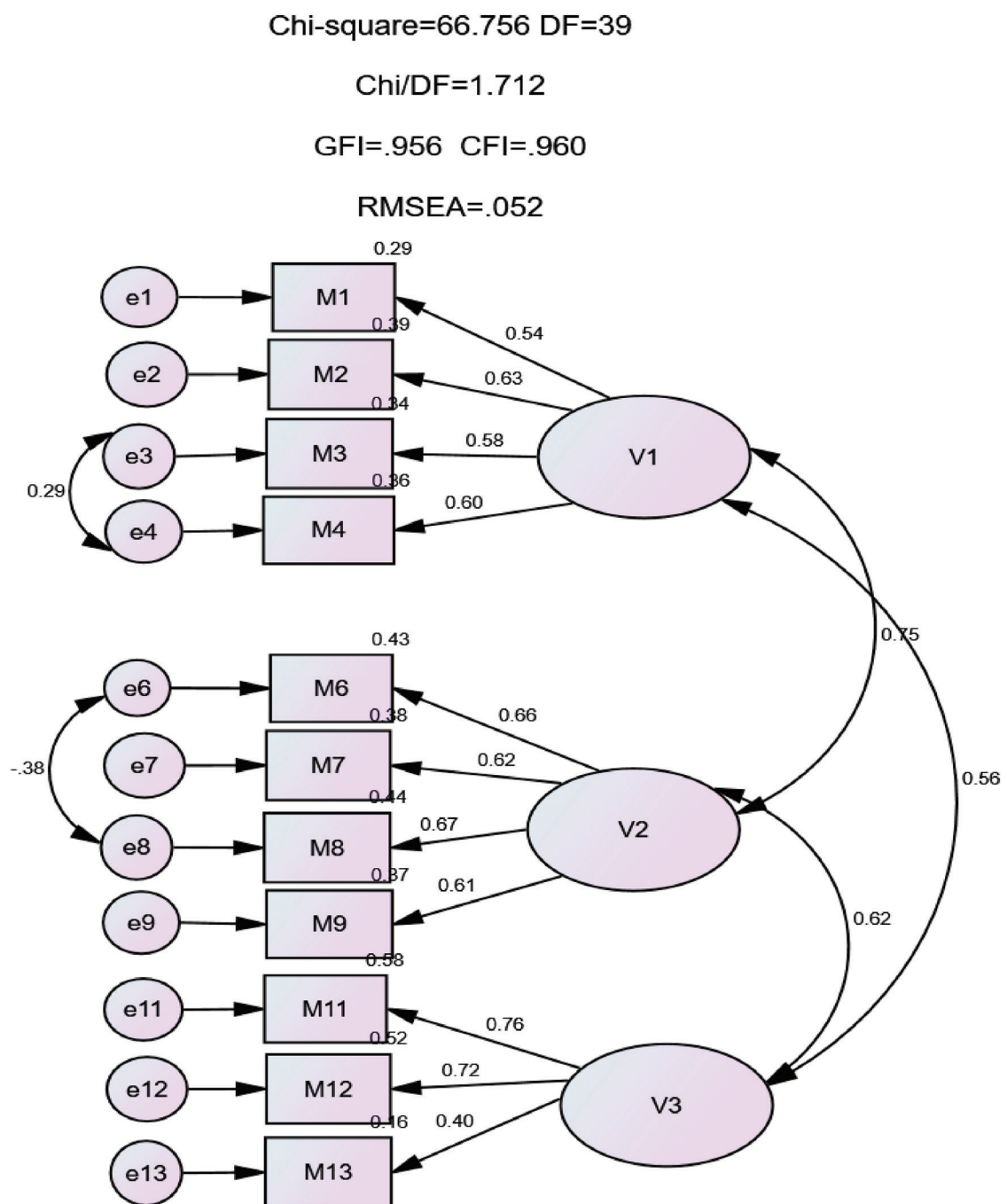


Figure 4 Proposed CFA model.

Table 2. Model fit of mediating factors

Model fit indices	Suggested	Obtained	References
RMSEA	< 0.06 (90% CI ≤ 0.06)	0.052	Hu & Bentler, 1999; Brown, 2015; Kline, 2016
SRMR	≤ 0.08	0.0489	
GFI	≥ 0.90	0.956	
CFI	≥ 0.90	0.960	
Chi-square/df ratio	≤ 3	1.712	

**Figure 5** CFA model of mediating factors.

RQ 3: How does the interplay between test-takers' interpretation of the test design and their reasons for undertaking a preparation course relate to speaking scores?

To address Research question 3, structural equation modelling (SEM) was employed to test the hypothesised mechanism of washback on learning. The structural model aimed to test how the interplay between the factors directly from the target test, Interpretation of the test design (V1) and the mediating factors, including Self-perceived language proficiency (V2), Academic expectations (V3), and Individual differences (V4) can have a connection with the observed outcome variable, the Recent IELTS Speaking Test Score (RTS). The hypothesised structural model is visualised in Figure 5.

Following the procedures established for the prior analyses, data normality was confirmed via skewness and kurtosis tests, ensuring the dataset met the necessary assumptions for SEM. The SEM was estimated using Amos 29 (maximum likelihood estimation). The fit indices for the overall structural model (Table 3) met or exceeded the recommended thresholds, indicating that the hypothesised model provided a good fit to the observed data and lending structural validity to our adaptation of Bailey's (1996) washback model. However, the initial model fit assessment (Table 3) revealed factor loadings of 0.95 (I1) and 0.68 (I2), but the MCQ factor (self-perceived evaluation goals) had a low loading (0.13). Despite psychometric concerns, the factor was retained due to its theoretical importance in the study's framework, which emphasises test-takers' interpretations of test design as drivers of learning behaviours and scores. Removing it could undermine the model's theoretical validity.

Model fit indices met acceptable thresholds (Table 3). These indices are more directly aligned with the objectives of this study, which was the primary concern in validating Bailey's basic washback model in the Chinese EFL context of IELTS test preparation. To be more specific, this study primarily aims to validate the relationship between test-takers' interpretation of the IELTS Speaking test, their reasons for enrolling in preparation courses and the connection with IELTS speaking scores, within the framework of Bailey's (1996) Washback Model. In addition to the measurement model, the structural paths of the SEM analysis revealed the specific relationships between the latent constructs and the speaking score.

The correlated error terms introduced in the SEM (e6–e7, e9–e11, e16–e18) were theoretically justified rather than purely data-driven (Figure 6). Specifically, e6–e7 link items within self-perceived language proficiency that both address perceived speaking deficiencies, reflecting shared method variance and conceptual overlap (Fox & Cheng, 2007; Gan, 2009). The correlation between e9–e11 reflects closely related academic expectation items concerning score targets and institutional requirements, which prior studies show are often cognitively co-activated (Xie & Andrews, 2013; Dong, 2020). Similarly, e16–e18 connect individual difference items related to anxiety regulation and confidence building, constructs empirically shown to co-vary in test preparation contexts (Horwitz, 2010; Shih, 2007). These modifications preserve theoretical coherence while improving model representation. The results of the structural path analysis are presented in Table 4 and visualised in Figure 5 and Figure 6.

The results of the structural model indicated that none of the four latent washback factors demonstrated a statistically significant direct relationship with the IELTS speaking score (RTS) at the $p < 0.05$ level. In detail, V1 (Test Interpretation) exhibited the largest standardised coefficient among the positive predictors ($\beta = 0.108$), but this path was not significant ($p = 0.141$). V2 (self-perceived language proficiency) showed a non-significant, negative standardised relationship ($\beta = -0.164$). This suggests that the relationship between V2 and RTS is likely indirect, where low perceived proficiency drives the behaviour (test preparation course enrolment) rather than directly predicting the score outcome. V3 (Academic expectations) and V4 (Individual differences) also had weak, non-significant positive effects ($\beta = 0.096$ and $\beta = 0.087$), respectively.

Collectively, the factors accounted for approximately 3% of the variance in speaking scores ($R^2 = 0.030$). This finding, though limited, is consistent with prior research on washback (Xie & Andrews, 2013; Dong, 2020). Although these paths were not significant, the model's good overall fit suggests that the hypothesised constructs are meaningfully interrelated within a broader theoretical mechanism. This underscores the phenomenon's complex, non-linear, and multi-causal nature, suggesting that while the identified constructs are relevant, a significant portion of variance is attributable to other factors or more indirect pathways not captured in this model.

All in all, the data supported the theoretical structure of the hypothesised washback mechanism, in which test interpretation and mediating factors form a 'Process' component related to the 'Product' (the score). This is evidenced by the acceptable overall model fit indices (Table 3). However, the lack of significant direct effects indicates that this relationship is not simple or linear. While the model validates the interconnectedness of these constructs, their influence on the final score appears to be indirect, mediated by the specific learning behaviours and strategies adopted during test preparation.

Table 3. *Model fit of the washback on learning model*

Model fit indices	Suggested	Obtained	References
RMSEA	< 0.06 (90% CI ≤ 0.06)	0.056	Hu & Bentler, 1999; Brown, 2015; Kline, 2016
SRMR	≤ 0.08	0.0671	
GFI	≥ 0.90	0.922	
CFI	≥ 0.90	0.908	
Chi-square/df ratio	≤ 3	1.803	

Table 4. *Unstandardised and standardised regression weights for structural paths to RTS*

Path (Predictor → Outcome)	Unstandardised Estimate (γ)	S.E.	C.R.	p-value	Standardised Estimate (β)
V1 (Interpretation of the test design)→RTS	0.010	0.007	1.471	0.141	0.108
V2(Self-perceived language proficiency)→RTS	-0.128	0.114	-1.116	0.264	-0.164
V3(Academic expectations)→RTS	0.071	0.114	0.624	0.533	0.096
V4(Individual differences)→RTS	0.054	0.069	0.783	0.434	0.087

Qualitative Results

This study used semi-structured interviews with Chinese IELTS test-takers to explore the mechanism of washback on learning. The first author served as a single coder to develop the code scheme and analyse transcripts, a process critical to data analysis integrity (Creswell & Creswell, 2017). Using a consistent set of criteria, the coder systematically identified and categorised themes, for example, coding statements like "wanting to improve general speaking skills" under "academic expectations" to minimise misclassification.

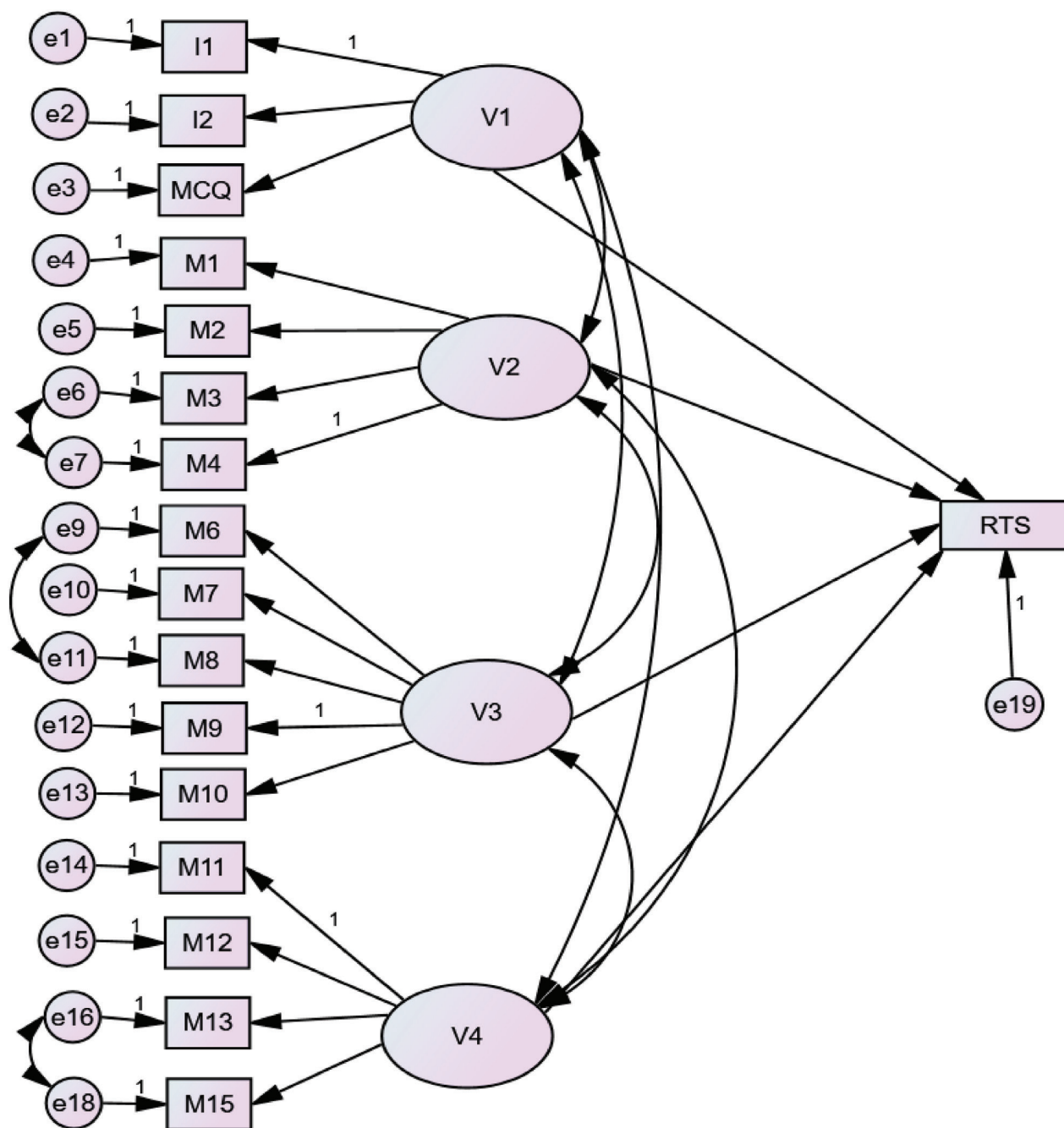
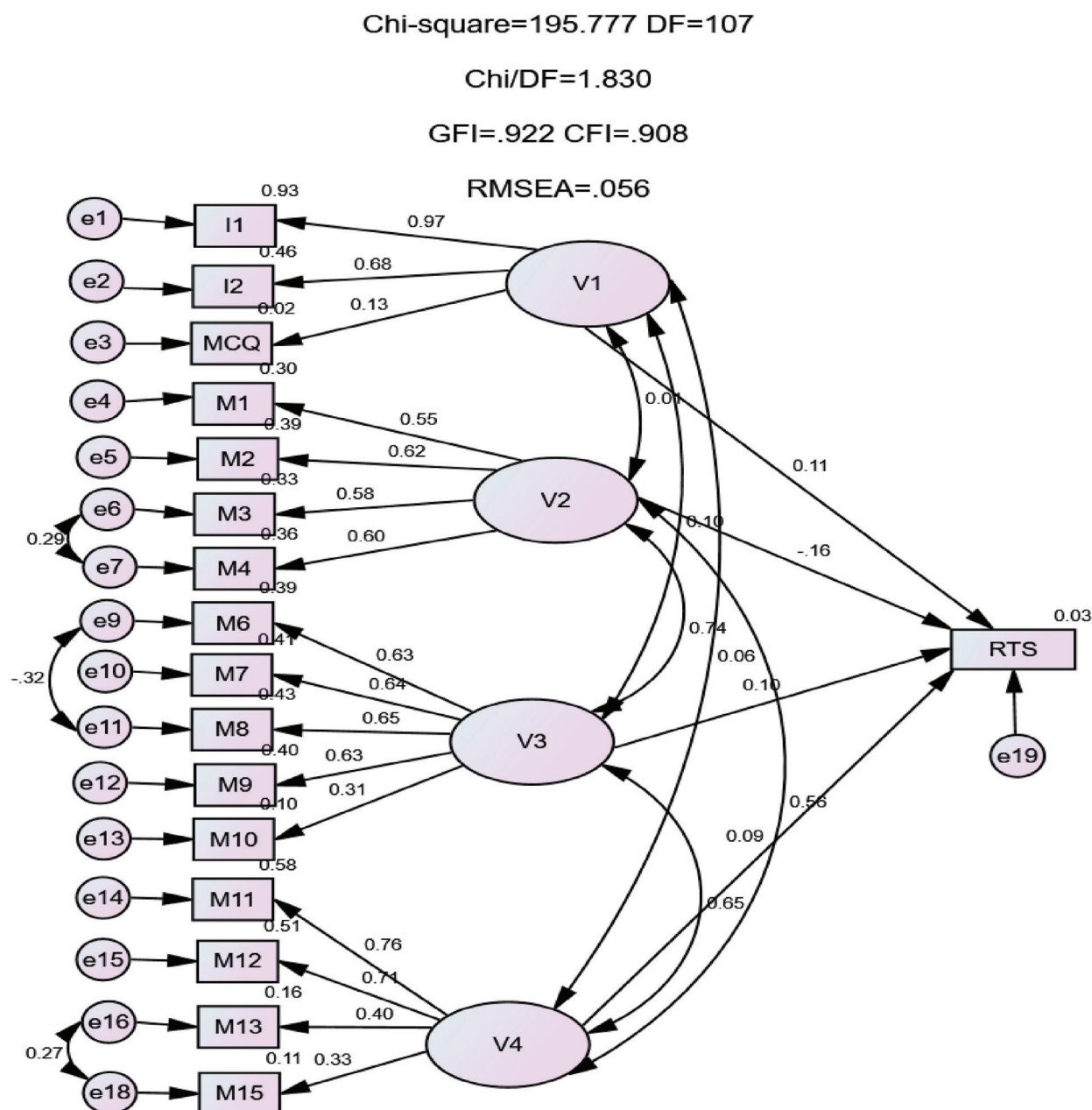


Figure 6 Washback on learning model.

The qualitative phase employed deductive thematic analysis guided by Braun and Clarke's (2006) six-step framework, focusing on test-takers' interpretation of the IELTS speaking test design and their motivations for enrolling in preparation courses. Nvivo 11 was used to organise transcripts, where the researcher first developed data codes to identify patterns relevant to the research questions. Codes were then applied to extract significant elements aligned with literature-derived themes (e.g., test design interpretation, mediating factors). Each theme was systematically described, and the final code scheme (see Appendix) was developed through thorough iterative refinement, ensuring alignment with both theoretical frameworks and empirical data.

The qualitative analysis, based on semi-structured interviews with 15 test-takers categorised by speaking scores (Group 1: ≥ 7.0 , $n = 3$; Group 2: 6.0–6.5, $n = 9$; Group 3: 5.0–5.5, $n = 3$), revealed distinct



patterns in test design interpretation (test construct, test criteria and self-perceived evaluation goals) and reasons for undertaking a test preparation course. The researchers did not use purposeful selection to invite participants with different band scores. The uneven distribution of the number of test-takers in each group was due to the typical score profile of Chinese IELTS test-takers. The mean score of test-takers in the dataset, along with the average score of Chinese IELTS test-takers from official statistics (IELTS Test Statistics, 2024), reveals that most test-takers achieved scores within the 5.5–6.5 range. This may account for unequal participant distributions and overly expansive middle groupings.

Test-takers' Interpretation of the Test Construct

Group 1 (≥ 7.0) associated the test with holistic academic skills. For example, Test-taker 1 (8.5 score) described Part 3 as requiring “logical argumentation similar to academic debates,” while Test-taker 2 (7.0) emphasised critical thinking in Part 3 tasks.

In Group 2, test-takers with scores of 6.5 (Test-takers 10 and 14) and 6.0 (Test-takers 3, 7, 8, 9, 11, 12) had various views. For Part 1, they generally saw it as evaluating basic communication and daily life interaction skills. In Part 2, they believed it was about constructing longer discourse, describing scenarios, or delivering impromptu speeches. For Part 3, they thought it evaluated advanced communication, critical thinking, and the ability to engage in academic like discussions. Although there were similarities, some test-takers had unique perspectives, such as test-taker 3 highlighting critical thinking and test-taker 10 emphasising quick responses.

All test-takers in Group 3 understood the IELTS speaking test as progressing from basic communication skills in Part 1 to more advanced language use in Part 2 and Part 3. They saw Part 1 as evaluating basic everyday conversation skills, Part 2 as assessing the ability to deliver monologues, and Part 3 as testing advanced communication and higher-order thinking skills. Test-taker 4 stated:

“I think one of the abilities this part (part 3) assessed was to answer questions in a timely manner, which is actually a bit like our PhD proposal defence. Or a slightly low-level version of the thesis defence, a slightly simplified version.”

In summary, the three groups of test-takers interpreted the IELTS speaking test differently based on their scores. Group 1 (≥ 7.0) viewed the test as assessing holistic academic skills, particularly in critical thinking and logical argumentation. Group 2 (6.0-6.5) saw the test as evaluating basic communication, discourse construction, and advanced communication in academic-like discussions. Group 3 (≤ 6.0) recognised a progression from a basic conversation in Part 1 to more advanced communication and higher-order thinking in Part 3. Overall, all groups noted a clear shift towards more complex skills in the later parts of the test.

Test-takers' Interpretation of the Test Criteria

In terms of the test criteria, in group 1, test-taker 1, with a high score, comprehensively understood the test criteria, including fluency, lexical resources, grammar, and pronunciation. Test-taker 2 had a broader understanding but lacked detail, and test-taker 6 did not explicitly express a detailed understanding. However, all in Group 1 were aware that the test evaluated a wide range of language skills.

Test-takers in Group 2 fragmented knowledge of the test criteria. Some, like test-taker 10 and test-taker 11, who did not clearly explain the criteria but recognised their importance. Test-taker 11 cited inconsistent examiner scoring standards in Macao vs. mainland China as a reason for not studying criteria deeply. Test taker 11 stated:

“In fact, I know little about it, and because I took the exam in Macao at that time. So, I heard it seems to be a little different from the mainland's scoring standards, so I did not go to see more of this aspect.”

Test-taker 14 remembered basic elements of the test criteria, but cannot recall all the key information. However, test-taker 9 and test-taker 12 had a solid grasp of the key elements of the test criteria, but test-taker 9 emphasised fluency and content, while test-taker 12 incorrectly believed pronunciation was less weighted in higher bands.

Test -takers in Group 3 had a general awareness of the importance of factors like sentence complexity and vocabulary range for scoring. However, none of them demonstrated a detailed understanding of the official four-dimension scoring criteria (fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation).

In short, Group 1 test-takers demonstrated a strong understanding of the IELTS speaking test criteria, with Test-taker 1 fully grasping fluency, lexical resources, grammar, and pronunciation, while others in the group had a broader but less detailed understanding. In contrast, Group 2 showed fragmented knowledge of the criteria, with some test-takers recognising their importance but struggling to explain them clearly, often due to external factors like varying scoring standards. Group 3 had a general awareness of key factors like sentence complexity and vocabulary range but lacked detailed knowledge of the four official scoring dimensions.

Test-takers' Self-perceived Evaluation Goals

Test-takers in Group 1 believed the IELTS speaking test aimed to assess comprehensive language proficiency, critical thinking, and analytical skills. They emphasised the importance of effective communication across different contexts and the evaluation of higher-order thinking skills, especially in Part 3.

Test-takers in Group 2, with a 6.5 score, believed the test aimed to evaluate proficient spoken English use, especially in academic settings. Those with a 6.0 score had diverse views, including evaluating practical communication skills, higher-order thinking, and the ability to organise thoughts and express opinions.

Test-takers in Group 3 recognised the test as a comprehensive assessment of speaking skills, from basic to complex. They believed success in the test required organising thoughts, presenting ideas effectively, and engaging in meaningful conversations. Some, like test-taker 15, also mentioned the role of luck in topic selection. Test-taker 15 stated:

“So, taking the IELTS test also has luck factors, because there are many topics, and you do not know which one you would encounter during the test. If you are very familiar with the topic, you can speak a lot and fluently. But if you encountered an unfamiliar topic, you did not even know a single word to express your ideas.”

All in all, test-takers in Group 1 viewed the IELTS speaking test as assessing overall language proficiency, critical thinking, and higher-order skills, especially in Part 3. Group 2, with scores of 6.0 and 6.5, saw the test as evaluating spoken English in academic settings, with some focusing on practical communication and organising thoughts. Group 3 recognised the test as a gradual assessment of speaking skills, from basic to complex, with some participants, like Test-taker 15, acknowledging that luck in topic selection could influence performance.

Reasons for Undertaking a Test Preparation Course

Group 1 test-takers commonly cited the need to familiarise themselves with the test format, comprehend the test requirements, practise particular skills, and learn test-taking strategies when enrolling in a preparation course. Test-taker 1 focused on getting familiar with the format, test-taker 2 on meeting test requirements and improving specific skills, and test-taker 6 on building confidence and acquiring strategies.

Group 2 test-takers frequently cited test anxiety, a desire to improve specific spoken skills, understanding the test structure, and meeting academic requirements as reasons for enrolling in a preparation course. For example, Test-taker 10's decision for taking the course was motivated by test anxiety and a desire to understand the test format. Test-taker 14 mentioned that understanding the test structure and improving skills were the key reasons. Test-taker 3 showed a preference for a structured learning environment and anxiety mitigation.

Reasons for Group 3 test-takers to enrol in a preparation course included achieving a desired score, understanding the test format and criteria, improving language proficiency, and reducing test anxiety. Test-taker 4 wanted to understand the test system and get a Band 6 score, test-taker 15 was motivated by self-perceived language limitations and anxiety, and test-taker 5 aimed for a desired score and valued the structured learning environment.

In terms of reasons, test-takers in Group 1 enrolled in preparation courses to familiarise themselves with the test format, improve specific skills, and develop test-taking strategies. Group 2 participants focused on reducing test anxiety, improving spoken skills, and understanding the test structure. Group 3 test-takers were motivated by goals such as achieving a specific score, understanding the test criteria, improving language proficiency, and managing test anxiety, often preferring a structured learning environment.

Discussion

RQ 1: What is Chinese IELTS test-takers' interpretation of the IELTS speaking test design?

The heterogeneous perceptions of the IELTS speaking test design (RQ1) reveal a critical gap between test-takers' interpretations and official constructs. Quantitatively, while mean scores suggested a moderate understanding of test structure, high standard deviations indicated widespread divergence in perceived assessed skills. This aligns with Chappell et al. (2019), who found that unclear skill expectations in high-stakes tests lead to ineffective preparation, particularly for productive skills.

Qualitative data further highlighted score-dependent interpretations: higher-scoring test-takers (Group 1) associated the test with holistic academic competencies (e.g., "academic debate"), while lower-scorers (Group 3) reduced it to basic communication. This mirrors Heidari Vincheh, Mirzaei and Roohani's (2024) finding that misaligned perceptions correlate with performance gaps. The persistent misinterpretation of evaluation goals underscores a systemic issue in Chinese EFL contexts, where exam-oriented culture often prioritises surface-level strategies over authentic language use (Hu & Trenkic, 2021).

This integration highlights a misalignment between intended and perceived test design, indicates a gap that threatens the test's consequential validity (Messick, 1996). While the IELTS speaking test aims to assess communicative competence in academic and real-world contexts, many learners instead prepare for surface-level performance, reflecting an exam-oriented interpretation of test design. Thus, test developers must improve clarity in communicating construct definitions. For example, the organisation of interactive workshops or visual rubric guides can be helpful to reduce such gaps, as advocated by some researchers (Xie & Andrews, 2013; Dong & Liu, 2022; Zhang & Bournot-Trites, 2021). Specifically, communicating with stakeholders (teachers and learners) to promote aligned and appropriate preparation practices is a crucial step for the testing body.

RQ 2: What mediating factors influence their decision to undertake a test preparation course?

The identification of self-perceived language proficiency, academic expectations, and individual differences as key mediating factors (RQ2) enriches washback research by empirically validating their combined impact on test-takers' decisions to enrol in preparation courses. While prior studies highlight direct washback from target tests in shaping learning behaviours and material selection, this study extends findings by demonstrating that mediating factors, often dismissed as construct-irrelevant (Messick, 1989) can significantly influence their learning behaviours. Quantitative analysis confirms these three factors are critical in explaining enrolment reasons, showing that Chinese IELTS test-takers' decisions are not

solely driven by test-specific elements but also by self-efficacy and external motivations, aligning with Green's (2013) framework.

Semi-structured interviews offer nuanced insights, which academic goals emerged as a universal mediating factor across proficiency levels, while qualitative data also highlighted reasons like skill improvement, strategy acquisition, and access to test materials, which were underrepresented in quantitative results. Individual differences (e.g., confidence, test anxiety, peer pressure) further shaped decisions, echoing Hawkey (2006) and Dong (2020).

The integration of these findings reinforces the idea that preparation decisions are shaped not only by test-related factors but also by individual psychological and social conditions, echoing research by Fox and Cheng (2007), Gan (2009), and Sato (2019). This finding also supports Shih's (2007) and Green's (2013) argument that internal and external learner characteristics transform test influence into specific learning behaviours. Thus, the "Process" component of Bailey's model is operationalised through the interaction between personal beliefs, expectations, and contextual pressures, determining whether washback is positive or negative.

RQ3: How does the interplay between test-takers' interpretation of the test design and their reasons for undertaking a preparation course relate to speaking scores?

The investigation of RQ3 uncovers a complex, multi-layered relationship between test design interpretation, reasons for enrolling in test preparation courses, and speaking scores, supported by both quantitative and qualitative evidence. The structural equation modelling (SEM) analysis revealed that the direct paths from the combined latent washback factors (V1-V4) to the final score (RTS) were statistically non-significant. The non-significant direct paths are critical as it does not undermine the model, but rather highlight the non-linear, multi-causal nature of washback (Xie & Andrews, 2013; Dong, 2020). It suggests that the primary influence of the identified constructs is on preparation behaviours (the Processes) rather than having a direct, measurable linear effect on the final Product (Score) in a cross-sectional model.

Qualitative analysis revealed the critical role of test-takers' interpretations of evaluation goals. Participants who viewed the test as assessing holistic communicative competence (Group 1) engaged in deep learning activities (e.g., analysing discourse structure, practicing impromptu responses), whereas those with narrow, score-focused goals (Groups 2–3) relied on surface-level strategies (e.g., memorising Part 2 scripts). This mirrors Cheng's (1998) findings on goal-driven learning behaviours, where misaligned perceptions of evaluation criteria lead to ineffective preparation.

The interplay between different factors and the connection with test scores was precisely articulated through synthesis of quantitative and qualitative findings. The non-significant direct link to the score suggests an indirect mechanism where the contextual mediating factors interact with test interpretation to determine the effectiveness of the learning process. For example, our qualitative data suggests that high test anxiety or low self-perceived proficiency may drive a narrow, test-focused interpretation of test criteria, leading to a reliance on surface-level, memorisation-based preparation strategies (an ineffective Process), which ultimately fails to translate into a higher score.

Integrating quantitative and qualitative data confirms the multifaceted mechanism of washback effect on learning. To be more specific, the mechanism of washback on learning is therefore better conceptualised as a cyclical and mediated process, rather than a simple causal chain. The findings also resonate with Xie and Andrews (2013) and Dong (2020), confirming that the predictive power of test-related constructs on test performance or test scores remains low due to the multiplicity of intervening learner

variables. In addition, this study's theoretical contribution lies in operationalising Bailey's (1996) "processes" component as the dynamic interplay between test design interpretation and mediating factors. Unlike prior models that isolate these elements, the framework highlights how interpretations of assessed skills and the interplay of individual mediating factors collectively have a connection with test scores. This reinforces Bailey's (1996) model by showing how participants (learners), processes (preparation behaviours), and products (test scores) are interconnected, even if the measured pathways are primarily indirect.

Implications and Limitations

This study extends Bailey's (1996) model by contextualising it within the Chinese EFL setting. It highlights how learners' interpretation and mediating factors collectively influence preparation behaviour and the connection with test scores. Thus, this study can provide implications to test developers and course designers.

For test developers, clearer communication of test construct and test criteria (e.g., visual rubrics, orientation sessions) may enhance test-takers' preparation behaviours and align with the target test goals. The collection of test-takers' views of the test can contribute to the ongoing evaluation of test validity, as Messick (1996) and Chalhoub-Deville and O'Sullivan (2020) argued. For course designers, they should change courses to move beyond memorisation and foster metacognitive strategies to help learners analyse scoring rubrics. These strategies promote stakeholder assessment literacy, which, as Chan and Luo (2021) noted, contributes to more equitable and informed participation in high-stakes testing.

There is no doubt this study has limitations. The generalisability of the study is limited by its use of convenience sampling. Stratified or multi-site sampling should be implemented in future research to incorporate underrepresented subgroups, for example, rural area learners or those preparing independently, in order to capture a wider range of views in terms of the washback effect on learning. Moreover, the cross-sectional design restricts causal inferences about how interpretation-behaviour linkages change over time. A longitudinal study may clarify the influence of consistent engagement with IELTS preparation course on both speaking skills and the advancement of self-regulated learning. Future studies may also examine the washback on learning mechanism across other skill sets (e.g., writing, listening) or various testing methodologies, providing comparative data on how contextual factors moderate the washback-learning dynamic.

Declaration of Generative AI Use

In preparing this manuscript, the authors employed DeepSeek (V3.2) for the specific purpose of enhancing text clarity through language polishing and paraphrasing. The core research contributions, including conceptualisation, analysis, and scholarly interpretation remain solely attributable to the authors, who reviewed and validated all content.

References

- Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14(2), 115–129. <https://doi.org/10.1093/applin/14.2.115>
- 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/10.1177/026553229601300301>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.

- Chalhoub-Deville, M., & O'Sullivan, B. (2020). *Validity: Theoretical development and integrated arguments*. Equinox.
- Chan, C. K. Y., & Luo, J. (2021). A four-dimensional conceptual framework for student assessment literacy in holistic competency development. *Assessment & Evaluation in Higher Education*, 46(3), 451–466. <https://doi.org/10.1080/02602938.2020.1777388>
- China Daily. (2025, November 3). *Overall average score for Chinese mainland IELTS test takers hits 6.5* [News website]. Retrieved February 16, 2026, from <https://www.chinadaily.com.cn/a/202511/03/WS69086704a310f215074b8aec.html>
- 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. (1998). Impact of a public English examination change on students' perceptions and attitudes toward their English learning. *Studies in Educational Evaluation*, 24(3), 279–301. [https://doi.org/10.1016/S0191-491X\(98\)00018-2](https://doi.org/10.1016/S0191-491X(98)00018-2)
- Cheng, L., & Curtis, A. (2004). Washback or backwash: A review of the impact of testing on teaching and learning. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing* (pp. 25–40). Routledge. <https://doi.org/10.4324/9781410609731-9>
- 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., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage Publications.
- 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, Article 100824. <https://doi.org/10.1016/j.stueduc.2019.100824>
- Dong, M., & Liu, X. (2022). Impact of learners' perceptions of a high-stakes test on their learning motivation and learning time allotment: A study on the washback mechanism. *Heliyon*, 8(12), e11910. <https://doi.org/10.1016/j.heliyon.2022.e11910>
- 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. <https://doi.org/10.1080/02188791.2021.1918057>
- Dong, M., & Fan, J. (2025). Asymmetry and stratification: Examining the washback effects of the National Matriculation English Test in Chinese high schools. *RELC Journal*. Advance online publication. <https://doi.org/10.1177/00336882251393153>
- Ferman, I. (2004). The washback of an EFL national oral matriculation test to teaching and learning. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 191–210). Routledge. <https://doi.org/10.4324/9781410609731-20>
- Freelon, D. G. (2010). ReCal: Intercoder reliability calculation as a web service. *International Journal of Internet Science*, 5(1), 20–33.
- 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. *RELC Journal*, 40(1), 23–41. <https://doi.org/10.1177/0033688208101449>
- Green, A. (2007). Washback to learning outcomes: A comparative study of IELTS preparation and university pre-session language courses. *Assessment in Education*, 14(1), 75–97. <https://doi.org/10.1080/09695940701272880>
- Green, A. (2013). Washback in language assessment. *International Journal of English Studies*, 13(2), 39–51.
- Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2010). *Multivariate data analysis* (7th ed.). Pearson.

- Hamp-Lyons, L. (1997). Washback, impact, and validity: Ethical concerns. *Language Testing*, 14(3), 295–303. <https://doi.org/10.1177/026553229701400304>
- Hamp-Lyons, L. (2000). Social, professional, and individual responsibility in language testing. *System*, 28(4), 579–591. [https://doi.org/10.1016/S0346-251X\(00\)00039-7](https://doi.org/10.1016/S0346-251X(00)00039-7)
- Hawkey, R. (2006). *Impact theory and practice*. Cambridge University Press.
- Heidari Vinchek, M., Mirzaei, A., & Roohani, A. (2024). A cognitive diagnostic approach to IELTS speaking test: Unveiling the subskills and test-takers' perceptions. *Language Testing in Asia*, 14(1), 11. <https://doi.org/10.1186/s40468-024-00311-2>
- Horwitz, E. K. (2010). Foreign and second language anxiety. *Language Teaching*, 43(2), 154–167. <https://doi.org/10.1017/S026144480999036X>
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 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>
- IELTS Partners. (n.d.). *Test statistics*. Retrieved February 11, 2024, from <https://ielts.org/researchers/our-research/test-statistics>
- IELTS Partners. (2025). *IELTS Academic format: Speaking*. <https://ielts.org/take-a-test/test-types/ielts-academic-test/ielts-academic-format-speaking>
- Iwashita, N., Sasaki, M., Stell, S., & Yucel, M. (2021). *Japanese stakeholders' perceptions of IELTS writing and speaking tests and their impact on communication and achievement* (IELTS Research Reports Online Series, No. 4). British Council, Cambridge Assessment English, and IDP: IELTS Australia. <https://www.ielts.org/teaching-and-research/research-reports>
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.
- 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>
- 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>
- Mizutani, S. (2009). *The mechanism of washback on teaching and learning* [Doctoral dissertation, University of Auckland]. ResearchSpace@Auckland.
- O'Connor, C., & Joffe, H. (2020). Intercode reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods*, 19. <https://doi.org/10.1177/1609406919899220>
- O'Kane, P., Smith, A., & Lerman, M. P. (2021). Building transparency and trustworthiness in inductive research through computer-aided qualitative data analysis software. *Organizational Research Methods*, 24(3), 456–487. <https://doi.org/10.1177/1094428120903216>
- Pearson, I. (1988). Tests as levers for change. In D. Chamberlain & R. J. Baumgardner (Eds.), *ESP in the classroom: Practice and evaluation* (pp. 98–107). Modern English Publications.
- Rahim, M., Sharifudin, M. A. S., Hamzah, F., Bidin, A., & Abd Hamid, N. A. (2024). Discrepancies between actual and perceived abilities in understanding argumentative texts amongst ESL students in Malaysia. *International Journal of Asian Social Science*, 14(8), 275–290. <https://doi.org/10.55493/5007.v14i8.5153>
- 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.
- 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>

- Spolsky, B. (1994). The examination-classroom backwash cycle: Some historical cases. In D. Nunan, R. Berry, & V. Berry (Eds.), *Bringing about change in language education* (pp. 55–66). University of Hong Kong.
- Tompkins, C. N. E., Burnley, A., & McNeill, A. (2021). Factors that influence smokers' and ex-smokers' use of IQOS: A qualitative study of IQOS users and ex-users in the UK. *Tobacco Control*, 30(1), 16–23. <https://doi.org/10.1136/tobaccocontrol-2020-056055>
- 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* [Master's thesis, University College London]. UCL Discovery.
- 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. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education: Vol. 7. Language testing and assessment* (pp. 291–302). Kluwer Academic.
- Wall, D. (2000). The impact of high-stakes testing on teaching and learning: Can this be predicted or controlled? *System*, 28(4), 499–509. [https://doi.org/10.1016/S0346-251X\(00\)00035-X](https://doi.org/10.1016/S0346-251X(00)00035-X)
- Wang, Y., & Huang, B. H. (2020). Washback of TOEFL preparation courses on students' attitudes and score improvement. *International Journal of Linguistics*, 12(3), 84–104. <https://doi.org/10.5296/ijl.v12i3.16940>
- 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>
- Xie, Q. (2015). Do component weighting and testing method affect time management and approaches to test preparation? A study on the washback mechanism. *System*, 50, 56–68. <https://doi.org/10.1016/j.system.2015.03.002>
- 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*, 2017(1), 1–59. <https://doi.org/10.1002/ets2.12192>
- Zhan, Y., & Andrews, S. (2014). Washback effects from a high-stakes examination on out-of-class English learning: insights from possible self-theories. *Assessment in Education: Principles, Policy & Practice*, 21(1), 71–89. <https://doi.org/10.1080/0969594X.2012.757546>
- 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/10.1177/0033688216631174>
- 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), 14. <https://doi.org/10.1186/s40468-024-00302-3>
- 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, Article 100977. <https://doi.org/10.1016/j.stueduc.2021.100977>
- 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(1), 5. <https://doi.org/10.1186/s40468-021-00211-x>
- Zhou, Y., & Yoshitomi, A. (2019). Test-taker perception of and test performance on computer-delivered speaking tests: The mediational role of test-taking motivation. *Language Testing in Asia*, 9(1), 1–19. <https://doi.org/10.1186/s40468-019-0086-7>

Appendix

Factors and observed variables of washback

Code	Factor	Observed variable
MCQ	V1 Interpretation of the test design	Multiple-choice questions of the interpretation of the IELTS speaking test construct
I1		Perceived intended washback of the test criteria
I2		Perceived intended washback of the test construct
M1	V2 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 would improve my speaking by learning from my classmates
M6	V3 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	V4 Individual differences	it can help me reduce my test anxiety and 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	IELTS speaking score	Recent IELTS speaking test score

Code Scheme of the interview

Main theme	Sub theme	Description	Example codes
Chinese IELTS test-takers' interpretation of the test design	Test-takers' interpretation of the skills the test aimed to assess them	Participants showed their own interpretation of the skills that the test assessed them in the three parts (Part 1, Part 2 and Part 3) of the IELTS speaking test. The interviewer asked the participants to provide their ideas of each part of skills the IELTS speaking test aims to assess.	1. Skills assessed in Part 1 2. Skills assessed in Part 2 3. Skills assessed in Part 3
	Test-takers' interpretation of the test criteria	Participants were asked about their understanding of the IELTS test criteria, especially to ask them about the familiarity with the test criteria. In addition, their thoughts of the importance for test-takers to understand test criteria were also asked.	1. Criterion Knowledge 2. Criterion importance in preparing for the test
	Test-takers' self-perceived evaluation goals	Participants were asked about their own belief in the skills the test assessed and the agreement with the skills that the official document claimed.	1. Their own belief of the skills assessed in the test 2. Alignment 3. Misalignment
Chinese IELTS test-takers' reasons for undertaking the test preparation course	self-perceived language proficiency	Participants indicated how they viewed their own language ability could affect their decision to take a test preparation course	1. Self-rated language level 2. How this affected their decision for undertaking a test preparation course
	academic expectations of taking the course	Participants expressed how learning goals and expectations of success affected their decision to take a test preparation course	1. Their academic goals when they decided to undertake a course 2. How the goals influenced their decision
	individual differences	Participants expressed how test anxiety and personal factors that could mediate test-takers' learning behaviours	1. Test anxiety 2. Personal factors (Teacher/Peer/Parent) 3. Other personal factors
	Other reasons	participants freely addressed any reasons that affected their decision to take the course.	1. Points or ideas that did not belong previous mentioned factors.

Ethical approval letter



Jawatankuasa Etika
Penyelidikan Manusia USM (JEPeM)
Human Research Ethics Committee USM (HREC)

Universiti Sains Malaysia
Kampus Kesihatan
16150 Kubang Kerian, Kelantan, Malaysia.
Tel. : +609 - 767 3000/2354/2362
Fax. : + 609 - 767 2351
Email : jepem@usm.my
Laman Web : www.jepem.kk.usm.my
www.usm.my

JEPeM Code :

Protocol Title: The Washback Effect of the IELTS Speaking Test on Chinese EFL Test-Takers

We wish to inform you that your study protocol has been reviewed and is hereby granted approval for implementation by the Jawatankuasa Etika Penyelidikan Manusia Universiti Sains Malaysia (JEPeM-USM). Your study has been assigned study protocol code **USM/JEPeM/PP/23030265**, which should be used for all communications to JEPeM-USM in relation to this study. This ethical approval is valid from **29th May 2023** until **28th May 2024**.

Study Site: China

The following document has been approved for use in the study.

1. Research Proposal

In addition to the abovementioned document, the following technical documents were included in the review on which this approval was based:

1. Participant Information Sheet and Consent Form (English version) – Questionnaire (Pilot study)
2. Participant Information Sheet and Consent Form (Chinese version) – Questionnaire (Pilot study)
3. Participant Information Sheet and Consent Form (English version) – Interview (Pilot study)
4. Participant Information Sheet and Consent Form (Chinese version) – Interview (Pilot study)
5. Participant Information Sheet and Consent Form (English version) – Questionnaire
6. Participant Information Sheet and Consent Form (Chinese version) – Questionnaire
7. Participant Information Sheet and Consent Form (English version) – Interview
8. Participant Information Sheet and Consent Form (Chinese version) – Interview
9. Questionnaire (English version)
10. Questionnaire (Chinese version)
11. Semi-structured Interview (English version)
12. Semi-structured Interview (Chinese version)

While the study is in progress, we request you to submit to us the following documents:

1. Application for renewal of ethical approval 45 days before the expiration date of this approval through submission of **JEPeM-USM FORM 3(B) 2022: Continuing Review Application Form**.
2. Any changes in the protocol, especially those that may adversely affect the safety of the participants during the conduct of the trial including changes in personnel, must be