Language Testing in Asia

RESEARCH Open Access

Face validity and washback effects of the shortened PTE Academic: insights from teachers in Mainland China

Jingwen Wang[1*], Ying Zheng[1] and Yi Zou[1]

*Correspondence:
Jingwen.Wang@soton.ac.uk

[1] Department of Languages, Cultures and Linguistics, Faculty of Humanities and Arts, University of Southampton, B65, Avenue Campus, Highfield Road, Southampton SO17 1BF, UK

Abstract

Pearson Test of English Academic (PTE Academic), a high-stakes English language proficiency test, underwent substantial revisions in 2021. The test duration was reduced from 3 h to 2 h by reducing specific task numbers and sections. This study investigates the impact of these changes on teachers' perceptions and teaching practices, areas previously underexplored in language assessment literature. It focuses on the implications of the test's modifications, examining the face validity and washback effects through the lens of teachers in mainland China. Semi-structured interviews with four experienced PTE Academic test trainers, who were familiar with both the original and revised test formats, reveals that the revised PTE Academic is perceived to maintain strong face validity, particularly noted in its academic authenticity, balanced skill structure, and perceived result accuracy. Additionally, most teachers perceived an increase in test difficulty. A mixed washback effect was observed: while improvements in students' language competence—a positive outcome—were noted, teachers also reported a continued reliance on teaching test-oriented strategies for higher-scoring tasks, indicating negative washback. This study highlights the significant implications of reducing PTE Academic's duration and offers targeted recommendations for its future improvement. These suggestions aim to enhance students' academic language skills, thereby better aligning PTE Academic with the practical language abilities required in university settings.

Keywords: PTE Academic, Test duration reduction, Face validity, Washback effects, Perceived difficulty, Teacher perspective

Introduction

Pearson Test of English Academic (PTE Academic), launched in 2009, is a globally recognised computer-based English language proficiency test. It is primarily designed to assess English competence in academic settings where English is the medium of instruction (Zheng & De Jong, 2011). Over the years, PTE Academic has increasingly become popular among international students for university applications in English-speaking countries, thus establishing itself as a widely-used tool for admission (Pae, 2012). In 2021, PTE Academic underwent significant revisions, reducing its duration from 3 h to 2 h by reducing specific task numbers and sections (Clesham, 2021). This trend towards

shorter test durations is also observed in other major English language proficiency tests such as the TOEFL iBT (ETS, 2023). These adaptations in the language testing field reflect a shift towards more efficient test designs, aiming to enhance accessibility while maintaining assessment quality.

Such high-stakes tests attract considerable attention from the research community, predominantly focusing on aspects of their validity and overall impact. Existing research has largely discussed the "pragmatic or statistical validity" (Mosier, 1947) of these tests, including construct, concurrent, or predictive validity (e.g. Dang & Dang, 2023; Gagen & Faez, 2024; Isaacs et al., 2023; Souzandehfar, 2024; Zheng & De Jong, 2011). However, face validity, which also measures test quality and assesses whether a test is subjectively viewed as covering the construct it is intended to measure, is often overlooked (Sato & Ikeda, 2015). Although some research has examined face validity from students' perspectives (e.g. Fan, 2014; Jackson, 2022; Sato & Ikeda, 2015; So, 2014), exploring teachers' views remains less common but is equally important.

Additionally, while the washback effects of these tests on teaching and learning are well-studied, they often concentrate on test preparation strategies, learner motivations, curriculum changes, and instructional adjustments (e.g. Alanezi & Alenezi, 2024; Chak, 2023; Gong, 2023; Gu, 2023; Sardi et al., 2022). The specific impact of changes in test length on teaching and learning, along with the relationship between face validity and washback effects, remains underexplored.

Recognising these gaps, the present study deliberately focuses on the perspective of teachers, as they are directly adapting their classroom teaching in light of these test modifications. This interaction provides them with unique insights into the practical and pedagogical implications of any changes to the test design. Their firsthand experiences are important for understanding how these adjustments affect teaching strategies and student learning outcomes—insights that are invaluable for refining the testing development process (Al-Adawi & Al-Balushi, 2015; So, 2014). Consequently, this study prioritises the perspectives of those most intimately involved with the application of the test modifications.

This study explores the revised PTE Academic as a case study. It aims to uncover how modifications to test length and question distribution influence teachers' perceptions of face validity, and the consequent washback effects on their teaching methods. Employing a qualitative methodology, the research involved semi-structured interviews with four experienced PTE Academic test trainers in mainland China. Selecting China as the context of this study was strategic, considering its role as one of the largest markets for PTE Academic. This choice aimed to ensure that the findings are representative of a significant segment of the global teaching community within this domain. The contribution of this study lies in its potential to inform test development and teaching practices, particularly in relation to how changes in test duration influence high-stakes language assessments.

Literature review

Face validity

Face validity was concisely defined as the "surface credibility or public acceptability" of a test (Ingram, 1977). It reflects the extent to which a test appears to assess the knowledge

or skills it claims to measure, according to the subjective judgement of an "untrained observer" (Davies et al., 1999, p.59). Anastasi (1982, p.136) offered a clearer explanation of who these observers might be, stating that face validity concerns whether the test "looks valid" not only to the examinees but also to administrative personnel who decide on its use, and other technically untrained observers. Essentially, face validity is about establishing a positive rapport and maintaining good public relations.

Building on these foundational ideas, recent discussions on face validity have increasingly centred on the alignment between test content and test administration settings. Holden (2010, p. 637) defined face validity as follows:

> *The appropriateness, sensibility, or relevance of the test and its items as they appear to those answering the test... More formally, it is the degree to which test respondents view the content of a test and its items as relevant to the context in which the test is being administered.*

This definition primarily considered test-takers' perception but failed to address a critical aspect: the intended users of a test, which should extend beyond just test-takers. Extending this further, Allen et al., (2023, p. 154) established a more comprehensive definition, arguing that face validity referred to "the clarity, relevance, difficulty, and sensitivity of a test to its intended audience".

Face validity is often overlooked in the fields of language testing and educational assessment, regarded by some researchers as irrelevant (Fan, 2014; Sato & Ikeda, 2015). Criticised for its reliance on the intuitive feelings of laypeople rather than a statistical model (Fan, 2014), face validity was frequently labelled unscientific and underestimated in its representation for "pragmatic or statistical validity" (Mosier, 1947) or "technical validity" (Stevenson, 1985). Stevenson (1985) further argued that incorporating face validity into the overall validity assessment of a test might overshadow more objective forms of validity, such as construct validity.

Despite these criticisms, some researchers have recognised the considerable importance that face validity holds. Nevo (1985) stated that face validity can significantly improve test-takers' motivation both before and during the test administration. Convincing arguments suggested that when test-takers perceived a test as face valid, they were more likely to prepare thoroughly and perform to their fullest potential (Alderson et al., 1995; Karelitz, 2013). Such enhancement in test-taking motivation emphasised the need for a deeper investigation into this aspect due to its potential influence on assessment outcomes (Xie, 2011). Teemant (2010) reported that test-takers might underperform if they did not perceive relevance between the test items and the intended constructs being assessed. This argument was later supported by Barnett et al. (2015). A lack of perceived alignment can result in the test not appearing to evaluate the intended skills, leading test takers to not fully engage with the test content in their preparation and actual test-taking, potentially resulting in scores that do not accurately reflect their abilities (Alderson et al., 1995; Bachman, 1990).

Apart from the impact on test outcomes, a limited number of empirical research projects have explored face validity from the perspective of stakeholders' perceptions of language test constructs. Notably, these studies have predominantly focused on viewpoints of test-takers. Brown (1993) investigated test-takers' reactions to a Japanese proficiency

test, discovering that participants reported low face validity. This perception was due to a misalignment between the test's specific objectives and the test-takers' existing skill sets, leading to dissatisfactions that could potentially influence their performance. Conversely, Thiel (1995) explored international students' perceptions of the IELTS test, finding it generally held high face validity, and that the test preparation was perceived as relevant to the target situation of their future academic studies.

More recently, Sato and Ikeda (2015) examined face validity by exploring how test-takers in Japan and Korea perceived the abilities being measured by high-stakes tests, especially whether these perceptions aligned with the intentions of the test developers. They discovered a moderately high agreement rate between the test-takers' perceptions and the developers' intentions. However, specific items intended to measure subtler skills, such as inferential understanding or implicit content, often did not align with test-taker perceptions, suggesting that such mismatches can lead to less effective learning outcomes and undermine the educational impact of the tests. Similarly, Rocha (2021) investigated undergraduates' perceptions of the TOEFL-ITP, confirming its face validity but noting the need for further research on time allocation due to some participants' dissatisfaction with the test's time limits. Zuhairoh et al. (2024) evaluated test-takers' perceptions of the final assessment test items for a 9th-grade English curriculum. Their findings revealed that the test's clear presentation and structured format not only enhanced its face validity but also increased its relevance and fairness by providing test-takers with straightforward instructions and a well-organised layout. In sum, these studies highlighted the importance of understanding test-takers' perceptions of face validity in developing fair and accessible tests. High face validity also appears to enhance test preparation and positively influence the long-term learning process.

Moreover, face validity might be considered more relevant when provided by individuals with subject expertise, such as test trainers in this case. However, research exploring face validity from the teachers' perspective is extremely rare. Al-Adawi and Al-Balushi (2015) investigated the face validity of an English language placement test used at their Colleges of Applied Sciences, discovering that teachers perceived the test as having low face validity. They recommended that future test designs include teachers' input. So (2014) conducted research that actively involved teachers in the development process of the TOEFL Junior® Comprehensive test, critically contributing to the final test design by integrating their insights and recommendations. These studies highlight the important role that teachers' insights play in shaping language assessments, improving their validity, and ensuring they align with the practical needs of educational settings.

Washback

It is well accepted that high-stakes tests significantly impact students; their future opportunities such as university applications may heavily depend on their test results. These tests also affect teachers, whose reputations and career progressions can be influenced by their success in preparing students to meet test requirements (Wall, 2012). In other words, when tests wield great power and have profound consequences for the lives of students, teachers, and other stakeholders (Cheng & Sultana, 2021), a phenomenon known as "washback" (Alderson & Wall, 1993; Cheng & Sultana, 2021; Cheng, 1997) may occur. Washback can be simply defined as the changes that testing brings to

classroom teaching and learning (Cheng & Sultana, 2021). Alderson and Wall (1993) described washback as students and teachers doing things they would not normally do because of the test. Messick (1996) extended this definition to describe how tests influence language learners and teachers to engage in activities that may not naturally occur but either promote or inhibit language learning. For instance, students may prioritise practising certain question types that carry more weight in scoring, or teachers focus more on teaching test-taking strategies than on language skills themselves, believing this could more rapidly improve test performance. Additionally, Morrow (1986) referred to this phenomenon as "washback validity", emphasising its important implications for the consequential aspect of construct validity.

The body of research on washback demonstrates its complex nature, indicating that it should not be perceived as a singular or straightforward phenomenon (Watanabe, 2004). Washback has extensive influences, affecting various aspects of teaching and learning (Cheng, 1997; Wall, 2012). Its emergence is shaped by a multitude of factors that dictate its manifestation and operation (Shohamy et al., 1996; Wall & Alderson, 1993). Researchers have categorised washback along several dimensions, including its specificity (general or specific), intensity (weak to strong), length (short-term or long-lasting), intentionality (unintended or intended), and value (positive or negative effects).

In response to the complexity of washback, researchers have developed theoretical models to better investigate its mechanisms. Alderson and Wall (1993) proposed 15 hypotheses related to washback from their "Sri Lanka study", with a primary focus on the micro-level aspects of teaching and learning influenced by tests. Complementing this, Hughes (1993, as cited in Cheng, 2004) introduced the "trichotomy backwash model" that centres on "participants", "processes", and "products". Building upon Hughes' (1993) model, Bailey (1996) more explicitly identified the "participants" as students, classroom teachers, administrators, and materials developers and publishers. Within this framework, the test may affect the perceptions and attitudes of these participants toward their roles. The term "process" refers to actions undertaken by these participants that could affect the learning process, while "products" refers to the learning outcomes and the quality of the learning achieved.

Washback is grounded in the belief that tests can and should drive teaching, which in turn influences learning (Cheng, 2004; Cheng & Sultana, 2021). Therefore, its initial impact is on the teachers' perceptions. Once these perceptions change, teachers might adjust various aspects of their teaching practices, including materials, content, and methodologies. These changes can subsequently spread to the entire teaching and learning process, ultimately influencing the learning outcomes. Over the past two decades, a surge of research has emerged, particularly within the English as a Second or Foreign Language (ESL/EFL) context, examining the complex nature of washback effects from language testing on teaching and learning (e.g. Alanezi & Alenezi, 2024; Chak, 2023; Cheng, 2004; Gong, 2023; Gu, 2023; Reynolds et al., 2018; Sardi et al., 2022; Xie, 2015). The consensus within the literature is clear: high-stakes tests, particularly international English proficiency exams important for admission, work, or migration decisions in English-speaking countries, have a considerable influence. This influence is particularly evident in how it shapes language teachers' perceptions and practices (e.g. Gu, 2023; Qi, 2005; Shohamy et al., 1996).

Sayyadi and Rezvani (2021) shed light on this aspect by exploring teachers' perceptions and teaching practices regarding oral questioning skills within the TOEFL iBT speaking test. Despite acknowledging the important role of questioning in academic interactions, instructors rarely emphasised students' questioning abilities or systematically planned and implemented classroom tasks to promote this skill, citing its exclusion from the test's format. They stressed that their primary obligation was to prepare applicants for a test with specific requirements and tasks, which did not assess proficiency in posing questions in English. Furthermore, instructors reported that their professional reputation and the pressure to achieve high test scores compelled them to concentrate on test-specific skills, at the expense of some essential academic skills.

Echoing prior findings, Homran and Asassfeh (2023) explored the washback effects of a Jordanian English exam from the perspective of teachers. Their study demonstrated that the high-stakes nature of the exam significantly shaped teaching practices, forcing teachers to emphasise specific content and skills such as reading, writing, and grammar, which directly influenced exam scores. Conversely, listening and speaking skills were often neglected, aligning teaching practices strictly with the exam's constructs and requirements. In addition, teachers frequently used previous exam papers as practice materials and integrated past exam items into quizzes and assignments to familiarise students with the question format and grading criteria. This phenomenon has similarly been observed by Puspitasari and Pelawi (2023) and Svantesson and Bahtiri (2024). Although extensive research has examined how high-stakes tests shape teachers' perceptions and teaching practices, the impact of other factors, such as changes in test duration, on these perceptions and practices is less well-understood.

Relationships between face validity and washback

The relationship between face validity and washback in language testing is complex and multi-dimensional. Face validity, as noted, refers to the extent to which a test appears to effectively measure its stated constructs to both test-takers and teachers, significantly influencing their behaviours, which is the essence of washback. When test-takers do not perceive an alignment between the test items and the intended constructs, the test may seem ineffective at assessing the intended abilities. Consequently, test-takers may not fully engage with the test content during their learning process, potentially leading to poor performance during the testing process (Alderson et al., 1995; Bachman, 1990). This disconnect results in negative washback (Cinkara & Tosun, 2017; Jackson, 2022; Sato & Ikeda, 2015), compromising the validity of the test scores and their interpretations.

Sato and Ikeda (2015) explored how test-takers' perceptions of the abilities assessed by items in a high-stakes English for Academic Purpose (EAP) test influenced their learning content. They concluded that effective washback might not be achieved if there was a gap between the test-takers' perceptions and the test committee's intentions. They also emphasised the importance of integrating test-taker feedback into test development to promote positive washback. In a related study, Cinkara and Tosun (2017) investigated the alignment between test-takers' perceptions and test developers' intentions in a small-scale university EFL program test. They observed low face validity, suggesting that

the mismatches in perceptions could lead to unintended washback effects on students' learning.

While existing research has examined how face validity influences test-takers, studies about its impact on teaching practices are less common. However, teachers' perceptions are equally critical and warrant attention in the literature. If teachers perceive a test as lacking face validity, they might adopt narrow test-taking strategies or emphasise less important proficiency skills. Such misalignment between teachers' focuses and the tests' intended goals can lead to instruction that fails to genuinely enhance students' proficiency. Jackson (2022) explored teachers' perceptions of the face validity of the General Aptitude Test. The results revealed that face validity substantially influenced teaching methods. He further emphasised that an understanding of test objectives and specifications could encourage teachers to focus on teaching English abilities directly, rather than merely engaging in "teach to the test", thereby potentially mitigating negative washback.

In addition, we have identified another gap in the literature concerning test segmentation and its impact on teaching practices. When tests are segmented to assess specific skills, the perceived importance of each section may disproportionately influence teaching priorities. For instance, as Homran and Asassfeh (2023) noted, if the reading and writing components of an exam were deemed more critical to success, teachers may allocate more instructional time to these areas at the expense of listening or speaking skills. Such imbalances in skills can impede comprehensive language development, raising concerns about the structure and communication of test objectives. Additionally, variations in how test sections or items are weighted may compromise the test's construct validity, affecting its ability to accurately measure intended constructs. In a word, if face validity is perceived as compromised, it may result in negative washback, where teaching practices prioritise test-taking strategies over holistic language learning.

To address the identified gaps, this study focuses on the following two research questions (RQs):

- RQ1. How do teachers perceive the face validity of the updated PTE Academic in terms of changes to test length and question distribution?
- RQ2. What specific washback effects have teachers observed in their teaching practices following the modifications to PTE Academic?

Methods

PTE Academic

Since November 2021, PTE Academic has reduced its test duration from 3 h to 2 h, and the total number of questions presented to test takers decreased from 70–82 to 52–64, although all item types from the previous version were retained. Detailed information about these changes can be found on the PTE official website (https://www.pearsonpte.com/articles/pte-academic-just-got-better). The major modifications to the test format are outlined in Table 1.

Participants

The participants in this study were four PTE Academic test trainers from Mainland China, all of whom are native Mandarin speakers. Mainland China was strategically

Table 1 Summary of reduced test items in PTE academic

Test part	Item type	Original format	Revised format
Overall	Total testing time	3 h	2 h
Part 1: Speaking and writing	Total time	77–93 min	54–67 min
	• Describe image	6–7 items	3–4 items
	• Retell lecture	3–4 items	1–2 items
	• Answer short questions	10–12 items	5–6 items
	• Summarise written text	2–3 items	1–2 items
Part 2: Reading	Total time	32–41 min	29–30 min
	• Multiple choice	4–6 items	2–4 items
Part 3: Listening	Total time	45–57 min	30–43 min
	• Summarise spoken text	2–3 items	1–2 items
	• Multiple choice	4–6 items	2–4 items
	• Highlight correct summary	2–3 items	1–2 items

selected as the study context due to its status as one of the major markets for PTE Academic, where recent test modifications are likely to exert considerable influence. This choice improves the potential applicability of the findings to broader global trends within the teaching community. The participants were selected using purposive sampling, a method well-suited for effectively narrowing the participant pool and targeting individuals based on specific research questions and their deep knowledge of the subject (Thomas, 2022). This approach facilitated the selection of teachers with extensive experience and familiarity with PTE Academic, ensuring they could provide valuable insights aligned with the study's objectives. Thus, it was essential for the selected participants to have extensive experience preparing students for PTE Academic. They should have knowledge of both the previous and updated versions of the test, or at a minimum, be familiar with the versions. In addition, they were expected to teach all the skills assessed by the test to provide a comprehensive view of their perceptions on the test's various parts and its overall construct.

The selected teachers, anonymised as T01 to T04, were engaged in teaching PTE Academic across three different educational and training institutions. The group consisted of two males and two females, maintaining an equal gender distribution. They had an average age of 27.25 years and an average of 4 years of teaching experience in PTE Academic. Three teachers instructed in all assessed skills, while one focused exclusively on speaking, although this participant also possesses experience in other skills. Three teachers held master's degrees in Applied Linguistics or TESOL from Chinese universities, and one received his degree from a university in an English-speaking country. Two of the teachers held teaching certificates, which varied in terms of educational levels. Moreover, the teachers reported that PTE candidates in China typically prepared for their tests through self-study. For those requiring additional support, private teaching was preferred. Consequently, the participants indicated that their classes were mainly delivered online in a one-to-one format, although they occasionally conducted classroom sessions with sizes ranging from small (fewer than five individuals) to medium (10 to 20 individuals). Table 2 provides a more detailed background of each participating teacher.

Table 2 Profile of the teachers

Teacher	Gender	Age	Experience years	Qualification	Taught skill(s)	Delivery model	Class size
T01	Male	29	6	BA, MA, University Teacher Cert	LRWS	Online	Small
T02	Female	25	2	BA, MA	S	Online and in person	Small
T03	Male	28	4.5	BA, MA	LRWS	Online and in person	Medium
T04	Female	27	4	BA, MA, High-school Teacher Cert	LRWS	Online	Medium

In the "taught skill(s)" column, *LRWS* stands for the following: *L* listening, *R* reading, *W* writing, *S* speaking

Instruments and data collection

Given the nature of the research questions, this study employed a qualitative method to thoroughly investigate participants' perceptions of the face validity and washback effects resulting from recent changes to PTE Academic. This approach is particularly valued for its capacity to elicit detailed insights into complex phenomena (Maxwell, 2008). Primary data were collected using in-depth, semi-structured interviews, an effective method for gathering experiential and perceptual data (Naz et al., 2022).

The interviews were structured into four distinct parts to systematically address the study's objectives. Part 1 (questions 1 to 5) gathered academic and professional background information from the teachers. Part 2 (questions 6 to 12) explored their perceptions of the new PTE Academic's purpose and format (i.e. face validity), focusing on evaluating the test's ability to fulfil its intended purposes, the impact of changes on skill balance, and the test's perceived accuracy after revision. Part 3 (questions 13 to 21) assessed the direct washback effects of the new PTE Academic on teachers' response, investigating how the test changes influenced their reactions, teaching methods and focus areas, and the immediate challenges posed by the new format. Part 4 (questions 22 to 25) concluded the interviews by gathering insights into the effects of the test modifications on students' test results and evaluating the competitiveness of PTE Academic against other standardised English tests, thereby broadening our understanding of the new PTE Academic.

Interviews were conducted in Chinese, the native language of the participants, to ensure clarity and elicit more insightful responses. Each interview was audio-recorded and ranged from 30 to 45 min, averaging 37 min in duration.

Data analysis

The interview data were transcribed verbatim, and a thematic analysis was conducted as suggested by Kiger and Varpio (2020). After multiple reviews of the data to get familiar with its depth and breadth, the first coder developed an initial coding template that captured all emergent codes from the interviews. This coding process was facilitated using NVivo 14. To verify the coding's credibility, a second coder independently coded the transcription of participant T01, achieving a reliability index of 0.94 with the initial coder. The preliminary coding results were then reviewed by the research team, with the majority being accepted. Discrepancies were resolved through discussion and

re-examination of the original data, leading to adjustments in some codes and themes. Codes were systematically categorised into main themes derived from the literature review that informed our interview questions. For codes that did not fit the established themes, new themes were created to ensure a comprehensive and authentic representation of the participants' perspectives.

Results

This section presents the results of interview data from participants. The "RQ1: Teachers' insights on the face validity of the new PTE academic" section addresses the first research question by exploring teachers' perceptions on the face validity of the updated PTE Academic. It examines the alignment of the test with its intended purpose of measuring test-takers' academic skills, the balance of assessed skills and overall construct, and the perceived accuracy of test results compared to the previous version. The "RQ2: Washback effects on teachers' perception and teaching practices" section responds to the second research question by investigating the washback effects of the new PTE Academic on teaching behaviours. This analysis focuses on teachers' initial reactions to the test modifications, the subsequent changes in their teaching content and methods, and the challenges they faced in adapting to the new format.

RQ1: Teachers' insights on the face validity of the new PTE academic

Intended purposes alignment

As a recognised authority in standardised English proficiency testing, PTE Academic has been rigorously evaluated for its construct validity (Riazi, 2013; Zheng & De Jong, 2011). As the recent revisions did not modify the item types, our study shifted focus from the face validity of individual item types to a broader assessment of whether PTE Academic effectively measures the academic skills test-takers will employ in university settings. This investigation centred on the test's perceived academic authenticity and its alignment with its core objective: to engage test-takers with interactive and integrative tasks that reflect the practical use of English in academic contexts, as the test aims to mirror the linguistic challenges encountered in academic environments (De Jong & Zheng, 2011).

Participants frequently noted that integrated skills tasks such as "Retell Lecture", "Write Essay", and "Summarise Spoken Text" accurately mirrored the authentic academic activities test-takers may encounter in university settings, thereby enhancing the test's academic legitimacy. Moreover, PTE Academic attempts to closely simulate real-life academic scenarios.

> *T03: PTE Academic is definitely trying hard to mimic scenarios students are likely to face in academic environments in the future. Tasks such as note-taking during lectures ("Retell Lecture") are very representative. The language it uses tend to be more academic and formal. This consistency is maintained across different versions of the test, whether it's the 2 h or 3 h format.*

However, participants voiced concerns regarding the test's comprehensive alignment with authentic academic demands expected in higher education settings. T01, who concurrently served as a part-time PTE trainer and a full-time university lecturer, drew

parallels between the academic rigour of his teaching environment and PTE Academic's scope. T03 shared similar doubts, questioning the extent to which PTE Academic could effectively fulfil students' future academic requirements.

> *T01: There's no doubt that PTE gives students a feel for what academic settings are like, covering topics and situations they'd find in a university campus. The vocabulary used is academic and spans a broad range of disciplines. However, when compared it to the academic proficiency that universities actually require, PTE, like other standardised tests, falls a bit short. A good example is the demands of formal academic writing – things like writing standards and the proper use of references. That's an area where PTE doesn't quite measure up.*

To bridge the perceived gap in the academic authenticity between the PTE test and actual academic settings, the participant teachers provided valuable suggestions. T01 emphasised the importance of "length" in academic settings and suggested incorporating more extensive reading materials into the reading part of PTE Academic, drawing comparisons with the formats of IELTS and TOEFL iBT. Simultaneously, T02 proposed revisions to the speaking and writing parts, stressing the necessity for a broader range of tasks that better assess students' communicative abilities and critical thinking.

> *T01: One standout feature of academic settings is the focus on [length]. Students are often required to read long texts, write detailed essays, and deliver comprehensive presentations. In this respect, PTE doesn't quite meet expectations. For example, its reading materials are quite brief, typically around 300 words. This could be an area for improvement, perhaps by incorporating longer reading sections similar to those found in IELTS and TOEFL.*

> *T02: I believe there's definitely room to improve the speaking part. By adding more subjective tasks and reducing some of the objective ones, we could get a fuller picture of students' communicative skills and their ability to think critically. Also, the writing part could use more flexibility. Instead of just sticking to argumentative essays, introducing tasks that require discussing different viewpoints could provide a more comprehensive insight into students' writing capabilities.*

Balance of test design

Regarding the balance of the four assessed skills in the updated PTE Academic, all the teachers expressed general satisfaction. T02 commended the revised distribution for being more methodologically sound, which he believed better supports the development of students' authentic language proficiency in test preparation. He attributed this improvement to the reduction in item numbers for tasks that previously relied on pre-prepared templates or certain test-taking strategies. In contrast, the remaining three teachers observed subtle differences in skill balance between the old and new versions.

Furthermore, T01 and T04 agreed that PTE Academic tended to prioritise listening and speaking skills, particularly speaking, while the writing and reading parts appeared to be less emphasised. The perceived reasons for this imbalance vary: T01 suggested that the reduced focus on writing stemmed from the constraints of machine scoring

technology, especially in areas that needed discourse analysis. On the other hand, T03 observed that the scoring mechanism of integrated items in PTE might distribute scores across different skills from one item, affecting the perceived importance of each skill.

Participants globally expressed satisfaction with the time allocations for each part of the updated PTE Academic. T01 and T02 stated that the reduction in test duration was appropriately aligned with the decreased number of items, which they considered reasonable. In addition, T02 reported that she usually conducted mock tests before the official exams to help students improve their time management skills. These participants noted no significant changes in the time allocations across the different parts of the revised PTE, while T03 and T04, perceived the time allocation to the reading part to be insufficient, which indirectly increased its difficulty.

> *T04: …Actually, I don't see much impact on the speaking part since responses are timed and move quickly to the next question. The writing task ('Writing Essay') remains the same with twenty minutes allocated to it. Big changes occurred in the listening and reading parts, though listening adjustments are timed separately. So, the most substantial adjustments are in the reading part. With fewer 'Multiple Choice' questions, the time feels more constrained, [indirectly increasing the difficulty]. This arises because it's tougher to manage nearly the same number of questions in a reduced timeframe. Specifically, the reading part was reduced only slightly, from 15–20 questions to 13–18 questions, with no reductions in key tasks like 'Fill in the Blanks' and 'Re-order Paragraph.'*

Perceived accuracy of test results

Having taught both previous and updated versions of PTE Academic, participants shared their insights on the improved accuracy of test results with the new version. They unanimously observed that recent updates have produced more accurate assessments. T01 believed this improvement, to be due, in part, to advances in PTE's automated scoring technology. Although the updates to the scoring system were not announced concurrently with the launch of the new PTE version, the official PTE website advertises that the scoring system's accuracy is enhanced by vast data inputs (https://www.pearsonpte.com/scoring): in 2020 alone, over 678,000 examiner responses were incorporated into the algorithm, with continual annual updates to refine the process. This extensive data integration substantially promotes the reliability of the test results.

> *T01: …I've always felt the earlier scoring system missed the mark. It seemed to favour those fast speakers, equating speed with fluency, and quite lenient on pronunciation. However, with PTE's current machine scoring system, as they've fed more and more data into their AI model over the years, I believe it's getting stricter and more on point…*

On the topic of the consistency between the result of PTE Academic and other standardised language proficiency tests, teachers pointed out the following:

> *T04: Different tests might have their own focus areas. Just because a student excels in PTE Academic doesn't mean they'll perform equally well in other tests. However, fundamentally, these tests share similarities, which suggests that if you're good at*

one, you're likely to do well in others too.

RQ2: Washback effects on teachers' perception and teaching practices

Teachers' immediate reaction to the test revision

All four teachers expressed positive attitudes when asked about their initial reactions to the reduction in PTE Academic's test duration and the reasons behind this revision. They unanimously welcomed the decision to shorten the test duration to a more manageable 2 h, viewing it as a move toward greater efficiency.

> *T01: I found it quite tough for students to sit through a 3 h test. The 2 h version better suits the fast pace of modern life. It feels more welcoming and efficient. This improvement benefits both PTE staff and candidates by saving time. Being able to complete listening, speaking, reading, and writing within 2 h, without needing to book a separate slot for the speaking test, makes it a great test format!*

> *T03: I was actually quite pleased with the change to a 2 h version because the previous 3 h test was too long. Many of my students mentioned feeling fatigued during the longer test. A lot of them have also taken IELTS and TOEFL and were initially daunted by the 3 h PTE, especially those less enthusiastic about English or less proficient. The shorter 2 h version with fewer questions has been more readily accepted by the students.*

In addition to their feelings on the test's shortened duration, the teachers also shared their perceptions on the difficulty level of the revised PTE Academic. Three teachers believed the difficulty had increased.

> *T02: My initial concern was that the speaking and writing parts might have become more challenging due to the reduction in questions that could be addressed using specific test techniques. This means that the test design now places more emphasis on questions assessing real language ability.*

Conversely, T03 assumed it had become easier, although no explicit reasons were provided for this opinion.

> *T03: Before the new version was officially launched, there was considerable debate the changes in its difficulty. Given that the 2 h exam features relatively fewer questions compared to the old version and other exams, I initially thought it might be a bit easier for students.*

The diversity in teachers' perceptions of test difficulty suggests potential variation in teaching methodologies. These perceptions were not merely a superficial reaction, they profoundly impacted their teaching practices. A detailed exploration of how these perceptions influenced their teaching content and methods will be discussed in the subsequent section.

Updates of the teaching practices against the new test format

The perceived increase in test difficulty was primarily attributed to the reduction of item types that previously allowed reliance on test-taking strategies and answering templates.

This modification has shifted the test emphasis towards items that require test-takers to rely more on their genuine English competencies. A majority of participants shared how this adjustment influenced their teaching practices. Only T01 reported that his course structure and teaching content remained largely the same, only introducing students to the changes in the test. In contrast, the other teachers adapted their instruction to emphasise item types with greater weight—those that cannot be effectively tackled only through memorised strategies and templates. This change in focus has indirectly improved students' language abilities in preparation for the test.

T02: I've shifted my teaching focus to what I call 'capability questions,' such as 'Read Aloud' and 'Repeat Sentence' in the speaking part, and 'Fill in the Blanks' and 'Reorder Paragraph' in the reading part. These items now carry more weight as they assess practical English skills applicable in real-world scenarios, rather than those based on test strategies or memorised templates. Once we cover the answering strategies for other question types, we rarely revisit them.

T04: Previously, the 'Describe Image' task had around 6 to 7 questions. Now it's about 3. We used to give students various answering templates, but with fewer questions now carrying less weight, relying on templates can cause student to hesitate, affect their fluency and scores. As a result, we now emphasise encouraging students to answer questions independently, which naturally helps develop their individual competence.

Nevertheless, the fundamental orientation of test preparation courses, which predominantly aim to achieve high scores rather than improve language proficiency, remains largely unchanged. This emphasis is understandable given the commercial objectives of private training institutions, which are designed to help students receive desired scores in a short period.

T02: While the new course design greatly improves student capabilities, it is still primarily exam-oriented and hasn't fundamentally changed.

T03: …I don't usually call myself a 'teacher', I prefer the term 'trainer'. Our institution, and trainers like me, mainly aim to help students quickly achieve high scores. While improving students' abilities is the ideal outcome, our priority is often on optimising their performance on tests.

In addition, the new PTE did increase anxiety among students and added to the workload for teachers in the initial phase following the revision. However, these effects were temporary, subsiding about a month after its official launch.

T01: Most of the feedback I get from students about the new PTE is positive now, nearly two years after the changes. But right after the updates, many students were really anxious. They were nervous, thinking the test might become much harder. Back then, I spent a lot of my time just trying to calm them down.

T03: Many of our students schedule their PTE exams around their plans to study

abroad. The November 2021 changes to the test caused a lot of distress for those who needed their results by December. Our team worked overtime, analysing the modifications to understand their implications fully. We even introduced several extra open classes to help address the students' concerns. Fortunately, since PTE generally requires a short preparation period, acceptance of the 2 h format became evident just a few weeks after the changes were implemented.

Challenges faced by teachers post-revision

The challenges in teaching following the PTE Academic modifications were not directly related to the changes in the test itself. T01, T03, and T04 identified a common issue: a shortage of authentic practice materials, particularly for listening. Their teaching resources primarily consisted of authentic PTE Academic test papers from past and official guidelines. To address these material shortages, T01 successfully sourced alternative resources, while T03 and T04 enhanced their teaching with self-designed materials. This reliance on materials from previous test papers stressed the nature of test preparation courses, which typically focus on "teaching to the test".

Discussion

This study investigated the face validity and washback effects of the shortened PTE Academic from the teachers' perspective. RQ1 examined whether the revised test design was perceived as capable of accurately evaluating test-takers' academic English skills required in university settings. The findings indicated that participants broadly recognised PTE Academic as having high face validity, noting satisfaction with the quality of the updated test. They observed that the new version effectively simulated academic scenarios through its item design. However, they identified a lack of academic authenticity in certain items and skills, which did not fully meet the complexity required for future academic challenges. Consequently, participants recommended extending the length of reading materials to mirror the extensive reading reflected in university courses and suggested adding more subjective item types while reducing those that depend on memorisable templates or specific test-taking strategies. These adjustments would provide a more thorough assessment of students' communicative competence and critical thinking skills in academic contexts.

We strongly advise test developers to consider these insights, because integrating teachers' input can significantly improve item design and distribution, as supported by Al-Adawi and Al-Balushi (2015), Sato & Ikeda, 2015, and So (2014). Moreover, participants confirmed that the new PTE Academic maintained a valid construct in its assessed skills, the same as the predecessor, and seemed to accurately reflect students' English proficiency in test results. Although this perceived subjective validity and accuracy cannot formally serve as validity evidence, they emphasised the test's strong face validity from the perspective of its test users.

To investigate the washback effects of the revised PTE Academic, RQ2 investigated changes in teachers' perceptions and their teaching practices. It uncovered both positive and negative washback effects. On the positive side, teachers adapted their teaching strategies to focus more on item types that require genuine language competence rather

than rote techniques or memorised templates. This shift indirectly promoted students' authentic language use, encouraging them to engage with tasks using their own skills. Consequently, the revised test format urged teachers to concentrate on areas requiring true English proficiency.

However, most teachers reported ongoing negative washback, consistent with findings from prior high-stakes testing research (e.g. Barnes, 2017; Puspitasari & Pelawi, 2023; Sayyadi & Rezvani, 2021). These studies indicated that such exams typically led teachers to focus on practicing test items and teaching test-taking strategies, rather than developing students' actual language skills. Additionally, teachers reported allocating more time to skills deemed crucial for test success, particularly listening and speaking, a trend also observed in Homran and Asassfeh (2023). Given that most students who enrol in the preparation courses primarily aim to achieve high scores rapidly, it is reasonable for trainers to maintain a focus on test-oriented approaches, in line with the commercial objectives of their institutions.

This study also attempted to illustrate the relationship between face validity and washback effects, demonstrating that face validity significantly influences teaching practices—a finding confirmed by Homran and Asassfeh (2023) and Jackson (2022). Our analysis revealed that robust face validity promotes positive washback. Our participants demonstrated a thorough understanding of the objectives of PTE Academic, the specific skills assessed by each item type, and the changes in item distribution following the test revision. This understanding allowed them to quickly adapt their teaching content and focuses to align with the new test format. As noted, the revision sought to decrease the number of items that could be answered through test-taking techniques, encouraging teachers to prioritise language competence in their instruction, which in turn positively enhanced learning outcomes. Additionally, the study observed that teachers mainly relied on past test papers and official guidelines for teaching materials to ensure students were thoroughly familiar with the test items and grading policy, a behaviour also reported by Homran and Asassfeh (2023). While this "teaching to the test" approach is often critiqued, it stresses the importance for test developers to supply comprehensive official practice materials and clear guidelines to facilitate straightforward and effective test preparation (Zuhairoh et al., 2024).

Conclusion

This study explored the face validity of the revised PTE Academic from teachers' perspectives and examined the associated washback effects on their teaching practices. The findings indicated that the strong face validity of the updated PTE Academic has encouraged teachers to improve students' language proficiency rather than only focusing on test-taking strategies. Despite this positive aspect, the fundamental nature of test preparation courses—primarily oriented towards "teaching to the test"—remained largely unchanged. This study further demonstrated that strong face validity clarifies test objectives for teachers, allowing them to tailor their instruction more effectively and, consequently, improve student learning outcomes.

Regarding implications, our findings shed light on the theoretical understanding of how changes in test duration affect its perceived validity and subsequent teaching adjustments. While the arguments of this study are credible, the relationship between

teachers' perceptions of the test's face validity and modifications in their teaching practice requires further empirical validation. From a practical perspective, the results emphasise the importance for test developers to consider the broader effects of test design changes on teaching and learning dynamics. Incorporating input from teachers and test-takers may benefit the test design. It is worth noting that the suggestions proposed by teachers to improve PTE Academic's test development in the present study should be evaluated and investigated further across diverse educational settings to confirm their effectiveness and generalisability.

Several limitations of this study warrant acknowledgment. In terms of the potential biases, concerns about negative washback might not be universally applicable, as the majority of participants were exclusively full-time test preparation trainers. Their emphases typically prioritised optimising test scores using test-oriented approaches over developing authentic language competence. Additionally, the limited number of participants constrained our ability to perform a formal face validity assessment (Allen et al., 2023; Nevo, 1985), leading us to rely on subjective impressions that may not fully capture the complexity of face validity.

To address these limitations, future research could expand the sample size, and employ a quantitative approach to more thoroughly explore face validity and its potential influence on teaching (e.g. Frantz & Holmgren, 2019). Further studies are also recommended to broaden the current understanding of how various test modifications affect different stakeholders' perspectives. These studies could examine the impacts of test item modifications on teaching quality and student learning outcomes, as well as including a broader range of stakeholders, such as test developers, programme administrators, and policy makers, into the observation (e.g. Bukh et al., 2022; Terasawa et al., 2024; Xu & Liu, 2018).

Abbreviations

EAP	English for Academic Purpose
EFL	English as a Foreign Language
ESL	English as a Second Language
PTE Academic	Pearson Test of English Academic

Acknowledgements
The authors express their sincere gratitude to those involved in participant recruitment and to all participants, whose collaboration was instrumental to the success of this research.

Authors' contributions
Jingwen Wang was responsible for conceptualising the study, developing the methodology, analysing the data, and drafting the original manuscript. Dr Ying Zheng was responsible for overseeing the project design and execution, coordinating resources, reviewing and finalising the article. Dr. Yi Zou assisted in data analysis and critically reviewed and edited the manuscript.

Funding
The present study was funded by Pearson PLC.

Availability of data and materials
The datasets generated and analysed during this study are confidential due to the sensitive nature of the information and participant privacy concerns. They are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate
This study was conducted in accordance with the ethical standards set forth by the University of Southampton Ethics Committee (ERGO Number: 82376). Informed consent was obtained from all participants involved in the study. Furthermore, all procedures involving human participants adhered to the ethical standards of both the institutional and national research committees.

Competing interests
The authors declare that they have no competing interests.

Received: 11 March 2024 Accepted: 13 July 2024
Published online: 07 August 2024

References

Al-Adawi, S. S. A., & Al-Balushi, A. A. K. (2015). Investigating content and face validity of English language placement test designed by colleges of applied sciences. *English Language Teaching, 9*(1), 107. https://doi.org/10.5539/elt.v9n1p107
Alanezi, M. A., & Alenezi, A. A. (2024). The impact of the IELTS writing test on postgraduate students in Kuwait. *International Journal of English Language Teaching*, *12*(2), 42–51. https://doi.org/10.37745/ijelt.13/vol12n24251
Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.
Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115–129. https://doi.org/10.1093/applin/14.2.115
Allen, M. S., Robson, D. A., & Iliescu, D. (2023). Face validity: A critical but ignored component of scale construction in psychological assessment. *European Journal of Psychological Assessment, 39*(3), 153–156. https://doi.org/10.1027/1015-5759/a000777
Anastasi, A. (1982). *Psychological testing* (5th ed.). Macmillan.
Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
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
Barnes, M. (2017). Washback: Exploring what constitutes "good" teaching practices. *Journal of English for Academic Purposes, 30*, 1–12. https://doi.org/10.1016/j.jeap.2017.10.003
Barnett, L. M., Ridgers, N. D., Zask, A., & Salmon, J. (2015). Face validity and reliability of a pictorial instrument for assessing fundamental movement skill perceived competence in young children. *Journal of Science and Medicine in Sport, 18*(1), 98–102. https://doi.org/10.1016/j.jsams.2013.12.004
Brown, A. (1993). The role of test-taker feedback in the test development process: Test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing, 10*(3), 277–301. https://doi.org/10.1177/026553229301000305
Bukh, P. N., Christensen, K. S., & Poulsen, M. L. (2022). Performance funding: Exam results, stakes, and washback in Danish schools. *Sage Open, 12*(1), https://doi.org/10.1177/21582440221082100
Chak, M. (2023). Washback in language learning strategies under high stakes language testing–A study of the Hong Kong secondary system. *rEFLections, 31*(1), 1–24. https://doi.org/10.61508/refl.v31i1.269539
Cheng, L., & Sultana, N. (2021). Washback: Looking backward and forward. In G. Fulcher & L. Harding (Eds.), *The Routledge Handbook of Language Testing* (2nd ed., pp. 136–152). Routledge.
Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and Education, 11*(1), 38–54. https://doi.org/10.1080/09500789708666717
Cheng, L. (2004). The washback effect of a public examination change on teachers' perceptions toward their classroom teaching. In L. Cheng & Y. Watanabe (Eds.), *Washback in language testing: Research contexts and methods* (pp. 169–192). Lawrence Erlbaum Associates Inc.
Cinkara, E., & Tosun, Ö. Ö. (2017). Face validity study of a small-scale test in a tertiary-level intensive EFL program. *Bartin University Journal of Faculty of Education, 6*(2), 395–410. https://doi.org/10.14686/buefad.281870
Clesham, R. (2021). *PTE Academic research summary of shortened test form.* [Research Report]. Retrieved June 18, 2024, from: https://assets.ctfassets.net/yqwtwibiobs4/482yXLVNKc9txHhfr9c9i0/773696173ee3587f31a2576dd2b29029/2021_PTE_Academic_Research_summary_of_Shortened_Test_Form.pdf
Dang, C. N., & Dang, T. N. Y. (2023). The predictive validity of the IELTS test and contribution of IELTS preparation courses to international students' subsequent academic study: Insights from Vietnamese international students in the UK. *RELC Journal, 54*(1), 84–98. https://doi.org/10.1177/0033688220985533
Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge University Press.
De Jong, J. H. A. L., & Zheng, Y. (2011). *Applying EALTA guidelines: A practical case study on Pearson test of English academic.* [Research Report]. Retrieved December 15, 2023, from: https://www.pearsonpte.com/ctf-assets/yqwtwibiobs4/3khKhcSMi5BBM61VjSaF3u/a5b8e85580cfe4c22456623e9649b99b/Applying_EALTA_Guidelines_____A_Practical_case_study_on_Pearson_Test_of_English_Academic_-_John_H.A.L..pdf
ETS. (2023). *TOEFL iBT® enhancements debuting July 2023. ETS news & insights.* [Press Releases]. Retrieved June 18, 2024, from: https://www.de.ets.org/news/press-releases/toefl-ibt-enhancements-debuting-july-2023.html
Fan, J. (2014). Chinese test takers' attitudes towards the versant English test: A mixed-methods approach. *Language Testing in Asia, 4*, 1–17. https://doi.org/10.1186/s40468-014-0006-9
Frantz, A., & Holmgren, K. (2019). The work stress questionnaire (WSQ)–Reliability and face validity among male workers. *BMC Public Health, 19*, 1–8. https://doi.org/10.1186/s12889-019-7940-5
Gagen, T., & Faez, F. (2024). The predictive validity of IELTS scores: A meta-analysis. *Higher Education Research & Development, 43*(4), 873–888. https://doi.org/10.1080/07294360.2023.2280700
Gong, K. (2023). Challenges and opportunities for spoken English learning and instruction brought by automated speech scoring in large-scale speaking tests: A mixed-method investigation into the washback of speech rater in TOEFL iBT. *Asian-Pacific Journal of Second and Foreign Language Education, 8*, 1–23. https://doi.org/10.1186/s40862-023-00197-2
Gu, K. (2023). Washback effects of IELTS test on teachers' adoption of teaching materials in the classroom in China. *International Journal on Social and Education Sciences, 5*(2), 381–392. https://doi.org/10.46328/ijonses.513

Holden, R. B. (2010). Face validity. In I. B. Weiner & W. E. Craighead (Eds.), *The Corsini Encyclopedia of Psychology* (2nd ed., pp. 637–638). Wiley.
Homran, M. M. A., & Asassfeh, S. M. M. (2023). EFL high-stakes exams: Are we leading teachers as language teachers or test teachers? *Journal of World Englishes and Educational Practices, 55*(3), 47–55. https://doi.org/10.32996/jweep.2023.5.3.4
Hughes, A. (1993). Washback and TOEFL 2000. *Unpublished manuscript, University of Reading*.
Ingram, E. (1977). Basic Concepts in Testing. In J. P. B. Allen & A. Davies (Eds.), *Testing and experimental methods* (pp. 11–37). Oxford University Press.
Isaacs, T., Hu, R., Trenkic, D., & Varga, J. (2023). Examining the predictive validity of the Duolingo English test: Evidence from a major UK University. *Language Testing, 40*(3), 748–770. https://doi.org/10.1177/02655322231158550
Jackson, S. (2022). Student and teacher perceptions of an English language test for university admission: Understanding its face validity. *Journal of Language and Culture, 41*(2), 170–220.
Karelitz, T. M. (2013). *Using public opinion to inform the validation of test scores.* [Research Report]. Retrieved June 18, 2024, from: https://www.nite.org.il/files/reports/abstracts/e387_abstract.pdf
Kiger, M. E., & Varpio, L. (2020). Thematic analysis of qualitative data: AMEE guide no. 131. *Medical Teacher, 42*(8), 846–854. https://doi.org/10.1080/0142159X.2020.1755030
Maxwell, J. A. (2008). Designing a qualitative study. In L. Bickman & D. J. Rog (Eds.), *The SAGE Handbook of Applied Social Research Methods* (2nd ed., pp. 214–253). Sage Publications.
Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241–256. https://doi.org/10.1177/026553229601300302
Morrow, K. (1986). The evaluation of tests of communicative performance. In M. Portal (Ed.), *Innovations in Language Testing* (pp.1–13). NFER–Nelson.
Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement, 7*(2), 191–205. https://doi.org/10.1177/001316444700700201
Naz, N., Gulab, F., & Aslam, M. (2022). Development of qualitative semi-structured interview guide for case study research. *Competitive Social Sciences Research Journal, 3*(2), 45–52.
Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement, 22*(4), 287–293. https://doi.org/10.1111/j.1745-3984.1985.tb01065.x
Pae, H. K. (2012). A psychometric measurement model for adult English language learners: Pearson test of English academic. *Educational Research and Evaluation, 18*(3), 211–229. https://doi.org/10.1080/13803611.2011.650921
Puspitasari, M., & Pelawi, M. A. (2023). The mapping of mediating negative washback of the national examination. *Eltin Journal: Journal of English Language Teaching in Indonesia, 11*(1), 87–98.
Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing, 22*(2), 142–173. https://doi.org/10.1191/0265532205lt300oa
Reynolds, B. L., Shih, Y. C., & Wu, W. H. (2018). Modelling Taiwanese adolescent learners' English vocabulary acquisition and retention: The washback effect of the college entrance examination center's reference word list. *English for Specific Purposes, 52*, 47–59. https://doi.org/10.1016/j.esp.2018.08.001
Riazi, M. (2013). Concurrent and predictive validity of Pearson Test of English Academic (PTE Academic). *Papers in Language Testing and Assessment, 2*(2), 1–27.
Rocha, R. V. S. D. (2021). *Face validity of the TOEFL-ITP according to UnB students.* [Course Paper]. University of Brasília. Retrieved June 20, 2024, from https://bdm.unb.br/handle/10483/37987
Sardi, A., Surahmat, Z., & Nur, S. (2022). The washback of intensive TOEFL training program (ITTP) on student's learning motivation. *ELS Journal on Interdisciplinary Studies in Humanities, 5*(4), 593–597. https://doi.org/10.34050/elsjish.v5i4.24570
Sato, T., & Ikeda, N. (2015). Test-taker perception of what test items measure: A potential impact of face validity on student learning. *Language Testing in Asia, 5*, 1–16. https://doi.org/10.1186/s40468-015-0019-z
Sayyadi, A., & Rezvani, R. (2021). Questioning in TOEFL iBT speaking test: A case of washback and construct underrepresentation. *Language Testing in Asia, 11*, 1–18. https://doi.org/10.1186/s40468-021-00137-2
Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing, 13*(3), 298–317. https://doi.org/10.1177/026553229601300305
So, Y. (2014). Are teacher perspectives useful? Incorporating EFL teacher feedback in the development of a large-scale international English test. *Language Assessment Quarterly, 11*(3), 283–303. https://doi.org/10.1080/15434303.2014.936936
Souzandehfar, M. (2024). New perspectives on IELTS authenticity: An evaluation of the speaking module. *International Journal of Language Testing, 14*(1). https://doi.org/10.22034/ijlt.2023.409599.1272
Stevenson, D. K. (1985). Authenticity, validity and a tea party. *Language Testing, 2*(1), 41–47. https://doi.org/10.1177/026553228500200105
Svantesson, M. L., & Bahtiri, A. (2024). "But the national test is something else": Teachers' perceptions of how English teaching practices and learning behaviours are impacted by the oral subtest of the national tests in ninth grade.[Course Paper]. Malmö University. Retrieved June 21, 2024, from: https://www.diva-portal.org/smash/get/diva2:1868862/FULLTEXT02.pdf
Teemant, A. (2010). ESL student perspectives on university classroom testing practices. *Journal of the Scholarship of Teaching and Learning, 10*(3), 89–105.
Terasawa, T., Sudo, S., Kajigaya, T., Aoyama, R., & Kubota, R. (2024). Slogans as a policy distractor: A case of 'washback' discourse in English language testing reforms in Japan. *Current Issues in Language Planning,* 1–24. https://doi.org/10.1080/14664208.2024.2355016
Thiel, T. (1995). An analysis of the evolution of the IELTS test and an investigation of its face validity. [Doctoral Dissertation]. University of Tasmania. Retrieved June 20, 2024, from: https://figshare.utas.edu.au/articles/thesis/An_analysis_of_the_evolution_of_the_IELTS_test_and_an_investigation_of_its_face_validity/23241953
Thomas, B. (2022). The role of purposive sampling technique as a tool for informal choices in a social sciences in research methods. *Just Agriculture, 2*(5), 1–8.

Wall, D. (2012). Washback. In G. Fulcher & F. Davidson (Eds.), *The Routledge Handbook of Language Testing* (1st ed., pp. 79–92). Routledge.
Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing, 10*(1), 41–69. https://doi.org/10.1177/026553229301000103
Watanabe, Y. (2004). Methodology in washback studies. In L. Cheng & Y. Watanabe (Eds.), *Washback in language testing: Research contexts and methods* (pp. 19–36). Lawrence Erlbaum Associates Inc.
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. (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
Xu, Q., & Liu, J. (2018). Stakeholders' perceptions of TEM and its washback. In Q. Xu & J. Liu (Eds.), *A study on the washback effects of the test for English majors (TEM): Implications for testing and teaching reforms* (pp. 107–155). Springer.
Zheng, Y., & De Jong, J. H. A. L. (2011). *Establishing construct and concurrent validity of Pearson test of English academic.* [Research Report]. Retrieved December 16, 2023, from: https://www.pearsonpte.com/ctf-assets/yqwtwibiobs4/1C2B6CF2Za4MJFzt417Zbt/702f8feddc65ddcd94e7f89d382ab84c/Establishing_Construct_and_Concurrent_Validity_of_Pearson_Test_of_English_Academic_-_Ying_Zheng.pdf
Zuhairoh, Z., & Syafa'ah, N., & Kurniati, D. (2024). Content and face validity analysis on 9th grade final test items for secondary school level. *Prominent Journal of English Studies, 7*(1), 21–28.

Publisher's Note