Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 15, No. 1; 2025, March

ISSN 2476-5880
https://www.ijlt.ir

A Case Study of Washback and Test Preparation of the New Version of PTE Academic

Yi Zou[1]*, Ying Zheng[2], Jingwen Wang[3]

ARTICLE INFO

Article History:
Received: April 2024
Accepted: August 2024

KEYWORDS
Chinese test taker perception
PTE Academic
Test preparation
Test Revision
Washback Effects

ABSTRACT

The Pearson Test of English Academic (PTE-A), a widely used high-stakes language proficiency test for university admissions and migration purposes, underwent a notable change from a three-hour to a two-hour version in November 2021. The implementation of the new version has prompted inquiries into the washback effects on various stakeholders. Focusing on a small sample of Chinese test takers (n=10), this paper explores washback effects following the revision of PTE-A and the complexity of test preparation through in-depth semi-structured interviews. The findings suggest a shorter test length is preferred and several different methods are adopted for test preparation which gives evidence to the positive washback. However, participants reported some confusion regarding certain test items, leading to the adoption of construct-irrelevant methods. This, in turn, may affect the face validity of PTE-A. While addressing the literature gap in this field, recommendations for improving the test design to better meet test takers' needs are provided.

1. Introduction

Standardized language tests have been used for multiple purposes, and globalization has led to a significant increase in the use of English language tests. The pervasive use of English in major socioeconomic sectors such as business, technology, and various academic settings has driven a surge in demand among international students seeking admission to universities where English is the primary medium of instruction. In this context, universities rely on standardized language tests to provide reliable evidence of incoming students' English ability, ensuring they are prepared for their academic programs. The new version of PTE-A has been designed to be useful for test takers in demonstrating their English language proficiency.

Launched globally in 2009, the PTE-A serves to measure candidates' ability to use English. In developing PTE-A, Pearson collaborated with Lexical Computing Ltd to establish the Pearson International Corpus of Academic English (PICAE). This corpus ensures the PTE-A accurately represents the English that test takers will need to understand and produce in academic settings where English is the language of instruction. The corpus comprises spoken and written data from five major English-speaking countries including curricular English found in lectures, seminars, textbooks and journal papers, and extracurricular English encountered in university administration and transcripts of radio broadcasts (see Ackermann et al., 2011). Besides its academic focus, it is noteworthy that the test is fully computer-based, with both test delivery and scoring automated by computer. Overall, PTE-A has become one of the leading large-scale proficiency tests affecting the lives of many international students, as high-stakes decisions often hinge on these test scores. The background of this study is that

[1] University of Southampton, Email: eyiechoyz@hotmail.com
[2] University of Southampton, Email: jingwen.wang@soton.ac.uk
[3] University of Southampton, Email: ying.zheng@soton.ac.uk

Cite this paper as: Zou, Y., Zheng, Y., & Jingwen, W. (2025). A Case Study of Washback and Test Preparation of the New Version of PTE Academic. *International Journal of Language Testing, 15*(1), 159–172. https://doi.org/10.22034/ijlt.2024.447992.1337

PTE-A was changed from a three-hour test to a two-hour version in November 2021, with a decrease in the number of test items while maintaining all item types. Compared to the IELTS, which typically takes around three hours to complete, the new two-hour version of PTE-A offers a compelling advantage in terms of efficiency (see Clesham, 2021; Clesham & Hughes, 2020). However, important questions arise about the validity and washback effects of the new version, which can include potential shifts in test-taker perceptions, experiences, and preparation strategies. It is increasingly recognized that test-taker perceptions represent a valuable source of face validity evidence and obtaining the positive and intended washback effects should be the responsibility of test developers (Sato & Ikeda, 2015). Thus, further understanding of the role of test-taker perception in relation to the test's face validity and washback is needed.

2. Review of Literature

2.1. Washback and Test Preparation

Washback refers to the influence of testing on teaching and learning. There has been enduring interest in the washback of high-stakes language tests on teaching and learning, where test preparation research is situated (e.g. Wall, 2012). These studies have been acknowledged as an integral element of validity inquiry (Cheng & Curtis, 2004). Alderson and Wall (1993) suggested that the term washback provides a useful metaphor to help us explore the role of language tests in teaching and learning. It allows for the possibility of the effects of tests on learning and teaching to be viewed as a continuum – stretching from negative at one end, through neutral, and into positive at the other end. Numerous theoretical frameworks have been developed to encompass the multifaceted factors involved in washback, with major washback models such as Hughes' (1993) trichotomy of washback: the *participants*, the *process*, and the *products*. Some empirical studies have found that test preparation activities are likely influenced by context-specific elements (e.g. Yu et al., 2017). These findings echo Messick's concern that test preparation research should take into account activities taking place in different contexts. The present study explores such context-related test factors by adopting components of Hughes' theoretical model and focusing on two components of the model: Chinese test takers (*participants*) and their test preparation activities (*processes*).

Specifically, Messick (1982) classified test preparation into three types. Type 1 test preparation improves test scores by improving construct-relevant aspects of language ability and therefore does not threaten score validity. Type 2 test preparation improves scores by reducing construct-irrelevant interferences in the test, such as unfamiliarity with the test tasks and test anxiety. These construct-irrelevant aspects of test performance may unduly lower scores. Type 2 test preparation is generally also seen as beneficial to score validity. Type 3 test preparation improves test scores by enhancing test-taking skills that are construct irrelevant. Such test-taking skills may help test takers to identify correct answers without possessing the relevant language ability, or without having to show the relevant language ability, and therefore may, if undetected, yield test scores that are inaccurately higher than they should be. Although the boundaries between these different types of test preparation activity can be difficult to discern, and the construct relevance of particular learning strategies may be hard to establish, Messick's framework remains a useful point of departure for analyzing the complex nature of test washback on test takers.

Although test takers may draw on a wide range of preparation activities, the majority of research studies examining test preparation have taken place in classroom settings, leaving self-access approaches largely unexamined. The term test preparation can be defined as the variety of activities undertaken to review the area of knowledge or skill sampled by a test. These activities may be undertaken both inside and outside the classroom. Effective test preparation is important when test outcomes have serious consequences for test takers and other score users. This is more likely to be the case with high-stakes English tests used for both university admission and migration like PTE-A, which is the focus of the current study. Test preparation has been classified by researchers in language education as part of the washback phenomenon, which concerns the influence of the test on language teaching and learning (Cheng & De Luca, 2011). The nature of test preparation may vary depending on a range of factors (Spratt, 2005), which include the test's content and format, the weightings of different components of the test, access to targeted preparation courses and resources, and whether test tasks measure language proficiency directly or indirectly via items that bear no obvious relationship to the

relevant language use context. Test preparation can be a double-edged sword. On the one hand, without any preparation, one may have doubts about the accuracy of the test score as a measure of ability, as a candidate's performance may be unfairly jeopardized by features of the test method with which they are unfamiliar. On the other hand, the wrong preparation, or too narrowly focused preparation, will either be counterproductive in the sense that it does nothing to improve language ability or might artificially boost the candidate's score for reasons of test wiseness unconnected with the ability the test is targeting (Ma & Cheng, 2018). These issues could affect the test's face validity, as they might undermine the credibility of test scores as a measure of the targeted ability.

The attainment of intended washback is a crucial part of test developers' responsibility. Language test designers are expected to carefully consider the intended impact on all the stakeholders in the initial phase of test development, a concept known as effect-driven testing (Fulcher & Davidson, 2007). Face validity is defined as "the degree to which a test appears to measure the knowledge or abilities it claims to measure, as judged by an untrained observer" (Davies et al., 1999, p. 59). Studies have found that low face validity can negatively affect test takers' performance; if the test is perceived as irrelevant to the claimed purposes, test takers may put less effort into preparation, resulting in scores that do not accurately reflect their ability (Brown & Abeywickrama, 2010; Kane, 2006). Further, Watanabe (2004) emphasizes that examining face validity from the test users' viewpoint is important to maintain test quality in terms of washback. Similarly, Brown (1993) asserts that test-taker perception can be used to develop a fair and accessible test. So (2014) argues that involving stakeholders in test development helps to improve test quality and fosters their acceptance of the test. In this context, the test takers' perceptions of the test constructs and skills measured in the test play a crucial role in achieving the intended washback.

In summary, this study seeks to explore the washback effects of the PTE-A test on test takers. The qualitative interviews focus on test-takers' perceptions of the shortened PTE-A test and their performance, accounting for factors such as learning behaviours, test-taking strategies, and motivation should also be explored. This study aims to extend the scope of existing literature, focusing on the revised PTE-A, to investigate whether the revision of a test, specifically the adjustment of test length, is a factor that impacts test takers' perceptions, leading to any changes in their learning practices. By examining these aspects, a more comprehensive understanding of the washback effects of the shortened PTE-A test on test-takers and their teachers can be gained. Such research can inform test design, test administration, and test preparation, ultimately contributing to improved language assessment practices and better educational outcomes for all stakeholders involved. The study is expected to provide Pearson and the related stakeholders with insights into achieving positive washback. By shedding light on the complex interplay between test length, test-taker perceptions, and learning practices, this study may help educators, policymakers, and test developers optimize test design and implementation, thus enhancing the overall effectiveness of high-stakes language assessments.

2.2. Relevant Studies

Washback research in the language testing field has focused on high-stakes English tests (e.g., Cheng et al., 2007). Test preparation, as one of the most immediate and observable impacts on teaching and learning, has emerged as a recurrent focal point in studies examining washback. Hayes and Read (2004), in their analysis of the IELTS academic module, found clear evidence of the washback effects of the tests. However, the outcomes did not seem to align with the anticipated positive effects when the focus of teachers and students narrowed to the practice of the test tasks themselves, rather than an improvement in language proficiency. A seminal investigation in this domain is Green's (2007) study, which delves into the influence of various types of preparation courses on IELTS writing scores. The findings of this study highlight the generally positive impact of preparation courses on test performance. Applying Hughes' (1993) trichotomy washback model, Rashidi and Javanmardi (2011) delineated washback effects on participants, the process, and the products. They found that IELTS preparation courses had positive impacts on the students' learning processes and their achievement in the examination. However, they also identified students' ambivalent expectations toward some aspects of these courses, which leads to no definite advantage of the preparation courses. These studies suggest that test preparation practices can generally have positive impacts on test performance but not necessarily help to improve their language proficiency, which may lead to weakening the predictive

validity of a test. While most previous studies have focused on correlations between test scores and students' actual academic success (Ihlenfeldt & Rios, 2023; O'Dwyer et al., 2018; Yen & Kuzma, 2009), some researchers have also paid attention to stakeholders' perceptions on the predictive validity. Esfandiari et al., (2018) designed questionnaires for TOEFL iBT candidates, teachers, and raters to explore the relationship between the performance on the TOEFL iBT and the target domain of English language use in an academic setting. In this study, there is a prevailing belief that students achieving higher scores on the TOEFL iBT also excel to a high standard in academic environments. However, the study acknowledges limitations in information about students' profiles, cautioning against generalizing one standardized test to predict all international students' academic skills in English.

Focusing on PTE-A, many researchers have paid attention to test takers' performance. For example, Zheng and Wei (2014) investigated Chinese and Indian test takers' performance on PTE-A and their motivation; anxiety-related factors and linguistic confidence. They found that Chinese and Indian students differ in their performance with observed score differences between test takers from the two countries, which can be partially explained by the variance in affective factors in their English learning and testing experiences. For the washback effects of PTE-A on learning, Ma (2014) focused on learners' test preparation activities in a specific test preparation center in China. Ma surveyed a group of 25 undergraduate students preparing for the PTE Academic at a commercial test preparation center in Beijing and looked at score gains between two live test sittings taken approximately six months apart. She identified a small score gain for listening and writing, reflecting a possible effect of test preparation on performance. Although her study included an interview component, she did not examine test preparation strategies adopted by students outside the classroom. Wei (2017a) conducted a study to investigate the relationship between the test taker's awareness of the constructs of integrated skills in PTE Academic and the learning strategies used in preparing for the test. He concluded that the test taker's selections of learning strategies can only be partially accounted for by their varying levels of understanding and awareness of the test. More specifically, there appeared to be a correlation between language learners' use of learning strategies, their awareness of the tested skills, and the types of audio inputs in the integrated skills tasks. Recently, Knoch et al. (2020) interviewed test takers who had taken the PTE Academic test more than once and analyzed their test preparation practices thematically along with their score gains in the four language skills. Among the four skills, speaking stood out as it was the language skill in which the largest score increases were identified; furthermore, many test takers reported making deliberate efforts to improve their speaking fluency and pronunciation, often based on advice from high-scoring friends or in coaching lessons, which suggested that speeding up their speech could lead to higher scores. Additionally, a number of test takers stated that practising particular speaking task types helped increase scores in speaking dramatically. Furthermore, in an attempt to understand how machine scoring works, many test takers also practiced through speaking into voice recognition systems, such as Google.

Moreover, as previously noted, the background of this study is that the test length of PTE-A has been shortened to two hours. Existing literature has addressed the impact of test length on test takers' performance. Some studies revealed that the increase in test length did not significantly impede test takers' performance despite a slight increase in reported fatigue (e.g., Liu et al, 2004). It appears that cognitive fatigue did not emerge as a significant factor influencing test takers' performance (e.g., Jensen et al., 2013). On the contrary, recently, Min and Bishop (2024) conducted a study to evaluate a large-scale multistage adaptive English language assessment. Their findings show that the test length for both the listening and reading tests could be shortened. With slight differences, the final ability estimates and reliability coefficients were comparable to those of the longer version of the test. Also, their study shows that the shortened version yielded slightly better measurement accuracy and efficiency.

While these studies focus on the factor of fatigue, it is imperative to consider other potential washback impacts, such as test-taker perception. Furthermore, test takers' perception of the fully computer-assisted test needs attention. Zhou and Yoshitomi (2019) investigated students' perceptions of computer-delivered speaking tests and their impact on test performance. The study explored whether students perceived their performance differently in a computer-delivered test compared to an in-person language assessment. While acknowledging concerns about technical difficulties beyond their control, most test takers did not perceive their performance as negatively affected by a computer-delivered test. As part of construct validity, in recent years, stakeholder perception has drawn increasing attention in

the field of language testing (Wei, 2017b; Xie & Andrews, 2013). Xie's (2013) study, for example, delves into the potential impact of test-takers' perceptions on the construct validity of CET-4. Despite potential misalignments between test-takers' perceptions of assessment demands and test developers' intentions, it is found that test takers adapt their preparation strategies based on perceived test areas of relevance.

In short, compared to studies on IELTS or TOEFL (e.g. Ariamanesh et al., 2023; Estaji & Banitalebi, 2022), there are limited studies on PTE-A. Further, to our knowledge, no research has given a special focus on the new version of PTE-A and its washback effects. Moreover, while existing studies have examined PTE test takers' performance and preparation strategies, the role of test-takers' perceptions is largely underexplored. Thus, to address the gap, this study explores the washback effects of the new version of PTE-A with insights from Chinese test takers and their test preparation practices.

3. Method

This study is part of the larger project on the face validity and washback of the new PTE-A version. The project consists of four datasets: quantitative analysis of the various psychometric parameters of the new version of PTE-A, qualitative interviews with test takers, qualitative interviews with test trainers, and content analysis of test items of the new version. This paper reports findings from qualitative interviews with Chinese test takers who have taken PTE-A for the purpose of studying in universities in the UK. The focus on Chinese test-takers' perception is initiated by an increasing number of Chinese students who choose to study in English-speaking countries or institutions where English is the medium of instruction. This highlights the necessity for studies on Chinese test-takers' perceptions and test-taking experiences.

Two main research questions were addressed in this study:

1. What are test-takers' perceptions of the new version of PTE-A following its recent changes?
2. How do test-takers' perceptions of the new version of PTE-A affect their test-preparation strategies?

3.1. Participants and Setting

All participants in this study were taking the PTE-A for admission to English-medium universities. Regarding their educational background, eight had completed their bachelor's degree at the time of the interview, with two of them already being accepted by universities in the UK for a master's degree. Another two participants held a master's degree. All participants reported having studied English for over 10 years. Four participants took the PTE-A new version once; eight participants had multiple attempts, with three of them having experience of taking the old three-hour version. While five of the participants were willing to provide their final score report, the other five were not for different personal reasons. All interviews were conducted in Chinese, the participants' first language, to ensure clarity and to help participants feel comfortable expressing themselves. However, the participants could choose to answer in English or in Chinese. With the participants' consent, the interviews were audio-recorded and subsequently transcribed verbatim for analysis. The extracts presented in this paper are translated by researchers of this study.

3.2. Instrumentation

This study adopted semi-structured interviews (Dörnyei, 2007). 10 participants were recruited randomly to assume that they represent the actual test taker population to a large extent. They were invited to engage in interviews from 35 minutes to one hour, depending on their availability and their willingness to participate. The interviews were conducted using interview guides (Appendix A) developed in English and translated into Chinese. Some adaptions to the interview guides were allowed in the actual interviews to refine interview questions to be more specific to each individual participant. A pilot study with one test taker was carried out to ensure the clearance and the validity of interview questions.

3.3. Procedures

The data source was compiled and thoroughly reviewed before data analysis. Thematic analysis was used for data analysis in this study. Following the literature review set out above, interview guides

were developed to explore the following themes: (1) their perception of the shortened version of PTE Academic, (2) their views on other large standardized English tests they are familiar with, for example, IELTS or Duolingo, with a focus on test item and test length comparing to PTE-A, (3) their test preparation practices. For those who prepared the test on their own, we paid attention to the learning resources that they used, in addition to the activities and strategies that they employed. For those attending preparation classes, we focused on their motivations and expectations from the lessons, the typical learning activities in class, and their perceived effectiveness of the test preparation classes in improving test scores and English ability. At the end of the interview, participants were provided an opportunity to share their perspectives on the strengths and weaknesses of the new version of PTE-A in focus as well as recommendations, along with any recommendations for future improvement.

The transcribed interview data was imported into NVivo 12. Both closed and open coding methods (Rivas, 2012) were employed to analyze the participants' responses to the questions. The closed coding method was chosen due to the semi-structured nature of the interviews in this study, where guided questions provided the structural framework for the coding scheme (Richards, 2014). Simultaneously, the open coding method was applied to identify other recurring themes emerging from participants' responses. Two researchers in the team coded part of the data independently, and then compared and discussed their coding results. Based on their discussions, the coding scheme underwent revision several times during the process. The findings were then interpreted and integrated to address the two research questions.

While this is a qualitative case study with a small sample of Chinese test takers, efforts were made to ensure the credibility, dependability, and confirmability. To ensure the credibility of the study, strategies including prolonged engagement with participants and peer debriefing were implemented. Researchers spent substantial time engaging with participants before and after interviews. Participants were invited to review the transcriptions and preliminary findings to ensure that their perspectives were accurately captured. The research team regularly engaged in discussions with colleagues to review and challenge the findings and interpretations. As for the dependability, two researchers independently coded the same data and then compared and discussed their results to ensure consistency and reliability in the coding scheme. To achieve confirmability, researchers constantly took account of their biases, assumptions, and reflections throughout the study, ensuring that their personal perspectives did not unduly influence the findings.

4. Results

4.1. Perceptions of PTE Academic Construct

4.1.1 Preference for the Two-hour Version

Regarding participants' perceptions of the PTE-A construct, the following discussion focuses on their perceptions of test length, content, and use. First, when comparing the three-hour version of the test to the two-hour version, test takers consistently expressed a preference for the latter. The primary reason for this preference was the perceived reduction in the test-taking effort required during the two-hour version. As reported by TT-07, for example, test takers could feel more fatigued when taking the longer test, which could influence their performance. Test takers report higher self-assessed stress if they took the three-hour version of the test. While the correlation between the test takers' actual performance and the test length needs further exploration, it is clear that participants in this study feel more motivated to prepare for the shortened version of the test: '*You will feel, it makes you feel more relax to prepare for the test, and less stressful*' (TT-08).

Extract 1

TT-07: The three-hour exam, for test-takers, it can indeed be quite *stressful* and a little *overwhelming*, especially for test like PTE that require you to maintain a high level of mental focus for an extended period of time. If the exam duration is three hours, it can significantly impact a person's overall performance, especially towards the end when they may become very *fatigued*. […] But if the test length is reduced from three hours to two hours, our performance can be much better. […]

Additionally, participants mentioned that the shortened version might be due to improvements in test design. In other words, it seems that participants hold the view that the shorter test length, the

better test design. Participants hold the view that if their English language competence can be measured properly with the two-hour version, there is no need to take the three-hour version, which is important for a test's measurement accuracy and efficiency.

Extract 2
TT-09: The reason could be that with the *continuous updates* in test items, test designers also noticed that certain questions needed to be retained. […] However, for certain items that do not contribute significantly to the overall score, such as some questions in reading part, *there may not be a need to include them anymore.* So, it can be seen as an *optimization*.

This finding provides valuable Chinese test takers' evaluation of PTE test designs, which can be considered by test designers to further optimize the designs. Besides the overall length of the test, test takers also give positive feedback on the duration of each individual section.

Extract 3
Interviewer: yeah, what do you think the time allocation should be for each section of the test?
TT-02: I think it's reasonable, each part.
Interviewer: Do you mean you just have enough time or more than enough?
TT-02: Just enough, just enough, you could say. For me, there's always *plenty of time* left for the last question.

Another interesting finding is that test takers also admit the authority of PTE Academic test design, saying '*I can't choose the test length, it's not what we can decide.*' (TT-08). While this comment shows test takers' trust towards PTE Academic authorities, it also reflects the necessity of researching the needs of test takers. In short, the data suggests that test takers generally hold a positive perception of the current test length and feel sufficiently motivated to prepare for it.

4.1.2 (Lack of) Comprehension of Test Content

As for the test content, the interview data suggests that test takers are generally aware of the integrated skills required in the test. However, there is also some confusion regarding the purposes of certain test items, which suggests their lack of comprehension of the test content. Participants specifically noted that some questions in the speaking section are designed just to get test takers to talk, without requiring meaningful content. For instance, in extract 4, TT-04 refers to questions that involve repeating sentences, summarizing a lecture, or describing the content of a chart or graph. She perceives these tasks as solely aimed at eliciting verbal responses rather than evaluating their understanding of meaningful content. This seems to suggest a gap between test takers' perceptions of the skills and abilities intended to be measured by the test and test designers' intended positive impact. However, it seems contradictory that all participants emphasized the importance of the speaking skill in the PTE Academic test. They considered speaking as the most crucial skill being tested and allocate considerable time and effort towards its preparation. Conversely, some participants perceived the reading section as challenging but comparatively less important in determining the overall score. As a result, they adopted strategies such as skipping certain parts of the reading section and allocating less preparation time. Clearly, participants' lack of comprehension of the test content is directly linked to their test preparation strategies, which will be further discussed in section 4.3.

Extract 4
TT-04: In terms of the speaking part, there are certain questions that are specifically designed to get you to start talking, *but they don't really have much meaning*. For example, there is a question to repeat a conversation or a sentence. It's like if there's a sentence played by a machine, and you listen to it, then you repeat the sentence. There is also a lecture where they will talk for about two to three minutes, and then they ask you to summarize what the lecture was about or something similar. Or when they give you a chart or graph and ask you to describe the content, I feel that they are solely designed to get you start talking *without much practical application*.

4.1.3 Motivations for Taking PTE Academic

Speaking of test use, participants in this study share the purpose of taking the PTE Academic test in order to meet academic institutions' English language requirement for admission. While this high level of motivation is identified, somewhat paradoxically, participants have mixed feelings about the extent to which their preparation for the PTE Academic test would be useful for improving their language abilities. The participants generally believe that preparing for PTE Academic can be beneficial for improving their listening and speaking skills, particularly in terms of communication in daily classroom interactions (see extract 5 for example). However, they also express the view that the test does not significantly contribute to their development of academic writing and reading skills. Notably, such mixed and ambivalent feelings are intertwined with their test preparation practices where construct-irrelevant methods are sometimes adopted.

Extract 5

TT07: For academic abilities, PTE focuses more on speaking and listening skills. It may help during our daily classroom interactions, but when it comes to writing academic papers and conducting literature reviews, I don't think it can be of help much. So, if we talk about its usefulness, I think it mostly helps in communication and listening skills.

4.2. Perceptions of PTE Academic vs Other English Test

In comparison to other language proficiency tests, most of the participants show a preference for choosing the PTE Academic due to their belief that it can help them achieve their goals within a shorter timeframe. However, when discussing the speaking assessments in both tests, most participants express a preference for the speaking test in IELTS. They believe that the IELTS places more emphasis on communication between individuals. Only one participant (TT-03) raises concerns about the subjectivity of the IELTS speaking test, mentioning that factors like the topic of discussion, accent, and test-takers' appearance could potentially influence the assessment. In contrast, the PTE Academic is considered more objective as it is entirely machine scoring.

Extract 6

TT-03: I feel that IELTS, as a human-based assessment, including the speaking test, has *a certain level of subjectivity***.** Factors, such as the topic you discuss, your accent, and even your appearance may influence the assessment. On the other hand, PTE is entirely AI-driven, *with machine scoring***.** I think that PTE requires a stronger focus on test-taking strategies.

Compared to Duolingo, in extract 7, TT-06 shares her experience of initially taking the PTE Academic test but later switching to Duolingo. She explains that her decision to switch was influenced by her struggles with meeting the standard of pronunciation in the PTE Academic test. Despite the switch, she still believes that the PTE Academic test better showcases test-takers' English proficiency.

Extract 7

Interviewer: Did you achieve the score you expected?
TT-06: No, I didn't. I switched to Duolingo later. My speaking skill was not good, and *I struggled with pronunciation*. Later, I found out that Duolingo has a scoring system with separate scores for listening, speaking, reading, and writing, along with subcategories. I couldn't overcome the pronunciation issue, so I decided to switch.
Interviewer: Compared to Duolingo, which test do you think better showcases your English proficiency?
TT-06: *Perhaps it's still PTE***,** but their emphasis is indeed different. Duolingo covers a broader range, but there are some parts that are similar to PTE. That is why I switched.

4.3. Test Preparation Practices

4.3.1 Test Preparation Classes

For test preparation practices, half of the participants reported that they prepared the test on their own, while the other half attended test preparation schools or classes. Commonly, their narratives suggest that they recognize the value of becoming familiar with the characteristics of each test item

through repetitive practice and exposure. All participants reported that they started by familiarizing the test including delivery mode, test formats, and specific test tasks. For those who experienced both self-preparation and preparation courses, preparation courses were viewed as more efficient to help them familiarize the PTE-A test. Additionally, as suggested in extract 8, the participants' distinction between test item types that require repetitive practice and those that require more strategic approaches showcases their awareness of adjusting preparation practices.

Extract 8
TT01: I usually start by dedicating two days to doing a large number of questions pool to familiarize myself with the characteristics of each question type. Some question types require a repetitive practice approach, where *continuous practice is necessary*. However, for question types that require more *strategic approaches*, I don't spend as much time on them. Instead, I focus on revisiting them a day or two before the exam.

According to the participants, the preparation classes were in either one tutor-on-one-student format or a group format. Some of these classes were delivered face to face, while others were streamed online. Those who attended the test preparation classes expressed hope that the tutors' assistance would help them become more familiar with the test with the assistance of the tutors and acquire test-taking skills to boost their scores. The participants also mentioned most preparation classes and schools can provide materials including authentic test items from previous PTE-A administrations, the skills and strategies to tackle each type of task, and some template responses. All of these materials aimed to boost students' scores on each task in the test. Overall, those who attended the test preparation classes spoke positively of them, believing that they were effective in helping them improve their scores to different extents. However, they also reported that they were much more geared towards improving test scores than improving their English ability in general. As indicated in Extract 8 below, the test taker perceives there is a gap between language tests and language use. From TT08's report, it seems meanings in language use are largely ignored among other emphases such as pronunciation and fluency while preparing for the test.

Extract 9
TT08: Actually, in PTE, I feel like you can score really high in speaking, but that doesn't necessarily mean your speaking skill is super strong. I believe *there's a bit of difference between the two* […] In PTE, they give you a passage, and you have to read it really well. At this point, I think the teachers focus more on correcting your *pronunciation, fluency, and overall speed*. I feel their emphasis is different. The teachers mostly cover how to read it smoothly and how to handle unfamiliar words and aspects like speed and intonation. This includes how to start at the beginning, how to read in the middle, how to end, and what to do when you encounter a stutter. It's more about these kinds of issues.

4.3.2 Self-preparation

As for those who prepared on their own, they referred to a wealth of test preparation resources that were available on the Internet or applications on their mobile devices. Specifically, the participants mentioned Alpaca Education and Firefly Enlightenment which were very popular among PTE-A test takers. Despite having access to a wide range of online resources, the participants reported their primary motivations for using these resources were to familiarize themselves with the task formats in PTE-A and to get more practice with authentic test items. Overall, most participants hold the view that the online resources were effective in helping with their test preparation. Extract 9 below illustrates this finding.

Extract 10
TT06: The first time I was preparing PTE test. *I felt lost* because preparing for PTE is different from IELTS. With IELTS, we all know there are listening, reading, writing, and speaking parts. But with PTE, even though it also includes listening, speaking, reading, and writing, there are many different question types within each section. […] So, the first step is to download the app, like Firefly

Enlightenment. Once you have the app, you can watch videos that explain the different question types. […] After that, you just need to start practicing on the app. *Just using your phone is enough*.

The interview data reveals that test-taker perceptions of the test are highly related to their access to resources, which shape washback to the test-takers.

5. Discussion

According to the findings of this study, participants consistently preferred the two-hour version of the PTE-A test over the three-hour version and showed increased motivation for test preparation. Participants expected better test performance with the two-hour version due to reduced fatigue. Besides, the test duration of each part of the two-hour version was perceived as appropriate. Additionally, participants held the belief that the shorter the test length, the better the test design. This finding gives evidence to the positive washback of the shortened version of PTE-A. However, the limited sample size of this study restricts the generalizability of this conclusion. Participants in this study expressed trust in the PTE-A test design without prioritizing their own needs. Nevertheless, researching test takers' needs could enhance positive washback effects and improve the test's face validity, as gaps may exist between the test designers' intentions and test takers' needs. These needs can include target language use domains, such as academic settings as mentioned by participants in this study.

Most participants chose the PTE-A test over other language proficiency tests (e.g., IELTS) due to the perceived benefits of meeting their goals (e.g., meeting university entrance requirements) more quickly with limited test preparation time. However, at the same time, they had mixed feelings about the extent to which their preparation for the PTE-A test would be useful for improving their language abilities. This is intertwined with the finding that there is a lack of comprehension of test content among test takers.

Participants reported their engagement in several different methods to prepare for the PTE-A test. These methods encompassed participation in preparatory courses, utilization of APPs, and the incorporation of online learning materials. Some test takers who took the test more than once engaged in a reflective assessment of their score reports each time. Combining their self-learning experiences with suggestions provided by preparatory tutors, they strategically adjusted their preparation efforts and time for subsequent test sitting. This finding contrasts with previous studies that have emphasized the test taker reliance on habitual test preparation methods and a reluctance to change the processes adopted (Zhan & Andrews, 2014). Test takers in this study show more agency in their test preparation. It should be pointed out that Zhan and Andrew focus on a different English test (CET-4) in China. Future comparison studies are expected to further explore test takers' agency in test preparation across different standard English tests.

In terms of speaking skills assessed in PTE Academic, test takers in this study reported that they were able to improve their scores by using construct-irrelevant methods. This included practices such as speaking continuously without pausing and changing their voice quality. Besides, test takers expressed some confusion regarding the purposes of certain test items, which led to a disparity between the test takers' perceptions of the skills and abilities to be assessed by the test and test developers' design. It seems, on the one hand, that the perceived importance and difficulty of PTE Academic has motivated test takers to make efforts in their preparation. On the other hand, it could be argued that test takers, while often falling into the trap of construct-irrelevant preparation activities in the short term, may come to realize over time that a focus on proficiency enhancement will serve them better in their quest for score improvement. For this reason, a careful examination of the preparation activities for this speaking test in future studies is expected to achieve a positive washback. For the other three skills, participants' report of score improvement in relation to test preparation is more closely linked to construct-relevant preparation, which could be taken as evidence supporting the validity of these types as measures of proficiency. Besides, participants believed that listening and speaking preparation practices were beneficial for them to engage in daily classroom discussions in their later academic study. However, they perceived limited contribution of writing and reading preparation practices to their academic study.

Although participants' test preparation practices were test-oriented to varying degrees, they expected some improvement in their language abilities through their preparation. Thus, this study calls

for the PTE-A test to strengthen their academic focus, particularly concerning the writing and reading test items. As suggested by the findings of the study, test takers engaged in some highly strategic test preparation practices. In addition, memorization of templates is also a common preparation practice among the test takers. These construct irrelevant (Messick, 1982) strategies can pose threats to the validity of the PTE-A test. Echoing Chen and Zheng (2022), the construct irrelevant strategies may be a concern for a test of academic English, as failing to write academically can lead to plagiarism in Higher Education assessment. PTE-A, being a fully automated test where the four skills are assessed by computer, claims reliability said to be assured by an algorithm. We suggest that Pearson, by making the scoring algorithms transparent to relevant test takers to some degree and clearly referencing language use in everyday contexts, can help test takers be better informed of the criteria and mechanisms against which their performances will be evaluated and with which they may align with their test preparation practices. This can improve test takers' expectations of how their preparations for PTE-A can transfer to their academic studies. We are also aware that Pearson may have reasons for not making these algorithms fully transparent, such as maintaining test security and integrity.

6. Conclusion

Overall, the current study provides generally supportive validity evidence for PTE-A. The findings add evidence to the explanation inference by shedding light on the skills, knowledge, and processes employed by test takers taking PTE-A. The finding concurs with Knoch et al. (2020) that it is necessary to consider the washback effects on new features which have been incorporated in language test design and how test takers' perceptions of a test impact the agency that they exercise in test preparation, which in turn shapes their preparation practices and potentially test results. The shortened test time is a significant change in PTE-A test design, which has washback effects on test takers' preparation practices. As suggested by Alderson and Wall (1993), washback is a dynamic and complex phenomenon. The washback effects on test takers' perspectives are challenging and complex, thus requiring examinations of various affective, cognitive, and social factors. It is important to note that the current study is based on a limited sample of Chinese test takers, as such the results may not necessarily generalize to other groups. While the findings of this study provide valuable insights into the washback of the new version of PTE-A, they should be interpreted with caution. Future studies are expected to consider different groups of test takers such as groups of different L1s and use different research methods such as quantitative or mixed-methods.

Acknowledgments

The authors would like to thank the participants of the study.

Declaration of Conflicting Interests

The authors declare that they have no conflicting interests.

Funding

This study is funded by Pearson.

References

Ackermann, K., De Jong, J., Kilgarriff, A., & Tugwell, D. (2011). The Pearson international corpus of academic English (PICAE). Proceedings of Corpus Linguistics. https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-47.pdf

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, *14*(2), 115-129. https://doi.org/https://doi.org/10.1093/applin/14.2.115

Ariamanesh, A. A., Barati, H., & Youhanaee, M. (2023). TOEFL iBT speaking subtest: The efficacy of preparation time on test-takers' performance. *International Journal of Language Testing*, *13*(2), 38-55. https://doi.org/10.22034/IJLT.2022.357001.1189

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*, 277-301. https://doi.org/https://doi.org/10.1177/026553229301000305

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (2nd ed.). Pearson Education.

Chen, Y., & Zheng, Y. (2022). A comparative study of Chinese test taker' writing performance in integrated and discrete tasks: Scores and recurrent word combinations in PTE Academic. In L. Hamp-Lyons & Y. Jin (Eds.), *Assessing the English Language Writing of Chinese Learners of English* (pp. 29-47). Springer. https://doi.org/https://doi.org/10.1007/978-3-030-92762-2_3

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: Research contexts and methods* (pp. 3-18). Lawrence Erlbaum Associates. https://doi.org/https://doi.org/10.4324/9781410609731-9

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

Cheng, L., Klinger, D. A., & Zheng, Y. (2007). The challenges of the Ontario Secondary School Literacy Test for second language students. *Language Testing*, *24*(2), 185-208. https://doi.org/https://doi.org/10.1177/0265532207076363

Clesham, R. (2021). PTE Academic research summary of shortened test form. Pearson. https://assets.ctfassets.net/yqwtwibiobs4/482yXLVNKc9txHhfr9c9i0/773696173ee3587f31a2576dd2b29029/2021_PTE_Academic_Research_summary_of_Shortened_Test_Form.pdf.

Clesham, R., & Hughes, S. (2020). Concordance Report PTE Academic and IELTS Academic. Pearson. https://www.pearsonpte.com/ctf-assets/yqwtwibiobs4/1hXHbkTLYCJly7JryACWjK/5a20dbe26d8ca2c36a3b0dd5a32868d7/2021_PTEA_2020_PTE_IELTS_Concordance_White_Paper.pdf.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge University Press.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford University Press.

Estaji, M., & Banitalebi, Z. (2022). Assessing test-taking strategies of IELTS test-takers: Development and validation of an IELTS test-taking strategy questionnaire. *International Journal of Language Testing*, *12*(1), 59-81. https://doi.org/10.22034/IJLT.2022.146984

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.

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

Hayes, B., & Read, J. (2004). IELTS test preparation in New Zealand: Preparing students for the IELTS academic module. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 97-112). Lawrence Erlbaum. https://doi.org/https://doi.org/10.4324/9781410609731-15

Hughes, A. (1993). Testing the ability to infer when reading in a second or foreign language. *Journal of English and Foreign Languages*, *10*(11), 13-20.

Ihlenfeldt, S. D., & Rios, J. A. (2023). A meta-analysis on the predictive validity of English language proficiency assessments for college admissions. *Language Testing*, *40*(2), 276-299. https://doi.org/https://doi.org/10.1177/02655322221112364

Jensen, J. L., Berry, D. A., & Kummer, T. A. (2013). Investigating the effects of exam length on performance and cognitive fatigue. *PloS one*, *8*(8), e70270. https://doi.org/https://doi.org/10.1371/journal.pone.0070270

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17-64). Greenwood Publishing.

Knoch, U., Huisman, A., Elder, C., Kong, X., & McKenna, A. (2020). Drawing on repeat test takers to study test preparation practices and their links to score gains. *Language Testing*, *37*(4), 550-572. https://doi.org/https://doi.org/10.1177/0265532220927407

Liu, J., Allspach, J. R., Feigenbaum, M., Oh, H. J., & Burton, N. (2004). A study of fatigue effects from the new SAT®. *ETS Research Report Series*, *2004*(2), 1-13. https://doi.org/https://doi.org/10.1002/j.2333-8504.2004.tb01973.x

Ma, J. (2017). *Examining Chinese test-takers' PTE Academic test preparation: Practices, effects and perceptions*. Pearson. https://www.pearsonpte.com/ctf-assets/yqwtwibiobs4/160GPz8klGBY6nO4fD2n1c/634df3001d1dff5e67a98b11dc4496b3/Examining_Chinese_Test_takers-_Jia_Ma.pdf

Ma, J., & Cheng, L. (2018). Preparing students to take test. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching*. John Wiley & Sons. https://doi.org/https://doi.org/10.1002/9781118784235.eelt0321

Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice. *Educational Psychologist*, *17*(2), 67-91. https://doi.org/https://doi.org/10.1080/00461528209529246

Min, S., & Bishop, K. (2024). A shortened test is feasible: Evaluating a large-scale multistage adaptive English language assessment. *Language Testing*, 1-22. https://doi.org/https://doi.org/10.1177/02655322231225426

O'Dwyer, J., Kantarcıoğlu, E., & Thomas, C. (2018). *An investigation of the predictive validity of the TOEFL iBT® test at an English-medium university in Turkey*. E*TS Research Report Series, 2018*(1), 1-13.

Rashidi, N., & Javanmardi, F. (2011). The IELTS preparation washback on learning and teaching outcomes. *Cross-Cultural Communication*, *7*(3), 132-144.

Richards, L. (2014). *Handling qualitative data: A practical guide*. Sage.

Rivas, C. (2012). Coding and analysing qualitative data. In C. Seale (Ed.), *Researching society and culture* (3rd ed., pp. 366-392). SAGE.

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/https://doi.org/10.1186/s40468-015-0019-z

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/https://doi.org/10.1080/15434303.2014.936936

Spratt, M. (2005). Washback and the classroom: The implications for teaching and learning of studies of washback from exams. *Language Teaching Research*, *9*(1), 5-29. https://doi.org/https://doi.org/10.1191/1362168805lr152oa

Wall, D. (2012). Washback. In G. Fulcher & F. Davidson (Eds.), *The routledge handbook of language testing* (pp. 79-92). Routledge. https://doi.org/https://doi.org/10.4324/9780203181287.ch5

Wei, W. (2017a). Can integrated skills tasks change students' learning strategies and materials? *The Language Learning Journal*, *45*(3), 336-351. https://doi.org/https://doi.org/10.1080/09571736.2014.905970

Wei, W. (2017b). A critical review of washback studies: Hypothesis and evidence. *Revisiting EFL assessment: Critical perspectives*, 49-67. https://doi.org/https://doi.org/10.1007/978-3-319-32601-6_4

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

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

Yen, D., & Kuzma, J. (2009). Higher IELTS score, higher academic performance? The validity of IELTS in predicting the academic performance of Chinese students. *Worcester Journal of Learning and Teaching*(3), 1-7.

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*, 1-59. https://doi.org/https://doi.org/10.1002/ets2.12145

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/https://doi.org/10.1080/0969594x.2012.757546

Zheng, Y., & Wei, W. (2014). Knowing the test takers: Investigating Chinese and Indian EFL/ESL students' performance on PTE Academic. *Asian EFL Journal*, *16*(1), 119-151.

Appendix A Semi-structured interviews

About PTE test-taking history

1. When did you take the PET Academic test, before Nov 2021 (the three-hour version), or after Nov 2021 (the two-hour version), or both?
2. If you have taken both, can you comment on the differences you noticed between these two versions?
3. How many times have you taken the PTE Academic test? Did you notice any changes or improvements in your performance across multiple attempts?
4. What motivated you to take the PTE Academic test initially?
5. How did you prepare for the PTE Academic test before taking it for the first time?

About the PTE two-hour version

6. How did you feel when you first heard about the new two-hour PTE Academic test? Did you experience any concerns or expectations?
7. Can you comment on the time allocated to answer the questions on PTE Academic? Do you think it is enough for each section?
8. How do you feel about the overall balance between the different sections of the new PTE Academic test? Do you think any particular section receives more or less emphasis?
9. What do you think PTE Academic is designed to measure, and do you believe the shortened test provides sufficient evidence for that claim? Could you comment on more specific skills like listening, reading, writing, and speaking?

About test preparation for the PTE two-hour version

10. Did you take any preparation courses for the PTE Academic test? How did these courses help you in your test preparation?
11. How have your study habits or strategies changed after you took the test preparation course? Do you think the change in your study habits/strategies has anything to do with the test being shortened?
12. In what ways has the shortened test affected your motivation to prepare for the PTE Academic test?

About test result and after-test reflection

13. What was your overall level of confidence in achieving your desired score on the new PTE Academic test? Have you achieved that goal? Would you like to comment on the four skills /item types in terms of difficulty or performance?
14. Do you believe PTE Academic is a good option to certify your English language level? Why or why not?
15. In your view, given the score you got from PTE Academic, do you think it provides a good indication of your academic skill? Why or why not?
16. What is/was your study plan? Are you/have you applied for a program in an English-speaking institution? If so, do you think PTE-A prepares you for the academic skills required for your further study?

About comparisons between the PTE test and other tests

17. Have you taken any other English tests besides PTE Academic? Can you please comment on your test-taking experiences in terms of test difficulty, item types, test design, etc.?
18. What item types do you think are useful in other tests but not in PTE Academic? Why do you think they are beneficial?
19. What do you think are the benefits and challenges of taking PTE-academic compared to other major international English tests?
20. Have you heard about TOEFL iBT being shortened to 2 hours? What do you think is the reason for that change, and how do you think it might impact test takers?