# Computer-based tests and machine marking: candidates' perceptions and beliefs about the test taking experience

**Authors: Mary Richardson, Sandra Leaton Gray and Jelena Popov, with Bryan Maddox**

## Introduction

When considering the present and future contexts of assessment research, it is important to explore the role of recent developments such as the increasing use of Artificial Intelligence (AI). Aspects of AI such as design and implementation, as well as its context within contemporary educational settings, are now necessary foci at both local and global levels. This research project presented a unique opportunity to build a detailed view of an AI-based English language testing model, PTE Academic, providing insights that hold the potential for transformational change. The research was carried out via a partnership between industry and academia and focused on the use of AI as embedded within an internationally recognised test. This meant that it also had an eye to contemporary global education contexts. As such, the enquiry builds on a long history of similar research in assessment, but with regard to new forms of 21st century examination practice and expectations.

The research team viewed this project as entering unchartered waters in the realm of educational assessment because the use of AI in this domain is still relatively limited (e.g., Richardson and Clesham, 2021). It therefore presented an opportunity to establish a foundation for creating a body of well-evidenced research interrogating the experiences of candidates. It has allowed us to learn more about how stakeholders understand and interact with a particular testing regime within the expanding parameters of AI in education, and as such, it acts as a case study of current practice with wider relevance to assessment research.

This study is an exploration of two domains from the perspectives of the candidates:

(a) documentation and explanation of their active experience of taking PTE tests; and

(b) their technical understanding and beliefs about how AI is used in PTE language testing.

Within the educational assessment sector, research about the use of AI in testing remains largely focused on the technical (Luckin, 2017), for example, accuracy in assessment scoring, the value of high stakes language testing, limitations in the use

of AI technology in testing, and practical issues such as rapid feedback and enhanced security (Chassignol *et al.*, 2018). The research findings reported here add to this growing body of literature by developing work on the personal experience of the candidate in an AI language testing setting. However, the focus of the research does not merely reflect the practical elements of test taking, the aim was also to collect data that helps to explore how candidates relate to their experiences and the ways in which they explain what it is like to be interacting with a test that employs AI technology. Putting the candidate at the heart of the discourse was guided by three research questions:

1. What is known about the role [their perception of planning for and taking the test] of the candidate in the AI-led language testing setting of PTE Academic?
2. How can we characterise the experience of candidates in AI-led language test domains?
3. To what extent can we use candidates' feedback to inform testing with AI components?

This qualitative research was not commissioned primarily as a validity study, however it did include data collection methods that align with validity-related themes. As might be expected, our focus on the candidates' experiences relates to the *Standards for Educational and Psychological Testing* (American Psychological Association *et al.*, 2014) proposition that three attributes underpin the efficacy of an assessment: validity, reliability, and fairness. The study can also be classified as interpretivist (Robson, 2002), in that the candidates are viewed as actors within a social world, and they are seen as understanding reality in complex and unique ways, with no search for a single definitive truth.

The initial scoping for the research began in 2020 and was interrupted by the Covid-19 pandemic during the first 18 months. However, from early 2021 until late 2022, the team was able to collect a rich data set that allowed us to establish an evidence base of candidates' experiences during the PTE Academic tests. This was achieved through an online survey and semi-structured interviews. Such an approach allowed for a good degree of triangulation of outcomes relating to current literature focused

on candidate experiences, perceptions of high stakes assessment, and views/beliefs about AI.

Three overarching categories were determined after the survey and interview data had been collected and analysed:

1. Test preparation
2. The test taking experience
3. Perceptions of AI and testing

As expected, a range of sub-categories are catalogued under these three headings, and it is here that we were able to analyse in depth some of the more detailed issues, views and characterisations of the test experiences. The next section is a short literature review that aligns with the key themes listed above and also considers some broader theoretical issues relating to conceptions of testing and test taking. These included issues of how candidates trust new technologies and their understanding of these form a part of assessment. In the third part of the report, we outline the results of the main online survey and interview results and consider the key themes emerging from the data. The concluding section returns to the research questions and presents questions and ideas for further exploration in relation to PTE Academic, both within the context of Pearson's work, and also for the further development of literature relating to candidate's views of educational testing that uses AI.

## Research in assessment and artificial intelligence

This section presents current research relevant to the PTE Academic test, locating that discussion within three broad themes of enquiry:

**1.** Test taking experiences (including preparation, test taking and perceptions),
**2.** Understanding and beliefs about artificial intelligence, and
**3.** Assessment validity and artificial intelligence.

A separate and fuller literature review was created as a part of this research. Here, we have selected appropriate literature to use as part of the discussion with the overall results, and we present it under the three headings listed above.

**Test taking experiences (including preparation, test taking and perceptions)**

The PTE Academic test is classified as 'high stakes' because it has significant reach and influence (Madaus, 1988). PTE results can determine candidates' routes into, and opportunities surrounding, tertiary education (Barkaoui, 2019) and/or university courses. They can even determine whether permission will be granted to work or reside in certain countries, such as Australia or Canada. Bennett (2015) argues that such tests are a key mechanism of social selection and sorting for hundreds of millions of people each year internationally, triaging access to different forms of education, professional advancement, and international mobility. The stakes (potential consequences) in the PTE Academic examinations are therefore likely to be higher than the informal, formative, embedded assessments that people may experience in other AI based language teaching and assessment. This is of course related to the fact that the results of PTE mean a great deal to candidates – success brings with it significant opportunities and failure can destroy their hopes. Such high personalisation of outcomes is a common reaction to high stakes tests, as discussed by Morris (2008) in relation to the reaction of students to law examinations, and Van Dijk et al in relation to testing in general (1999, 2002, 2003). It seems reasonable to assume therefore that the high stakes character of the PTE Academic test linked to emotionality may create specific perceptions and responses to the use of AI, which are not representative of wider uses of AI in assessment more generally.

Preparation (for example learning about likely test content, sitting mock examinations, and learning specific strategies for answering questions) is seen by candidates and educators as central to success in tests of any kind (O'Sullivan, Dunn and Berry, 2021). Given the significance of getting the right results in a language test such as PTE Academic, systematic and strategic planning (Razavipour, Habibollahi and Vahdat, 2021; Yu and Green, 2021) appears critical. As O'Sullivan et al (2021) argue, approaches to preparation for test-taking have long been recognised as challenging and contentious because success is not necessarily just determined by the quantity of revision. Additional social, cultural and economic factors that impact on the relative success of a candidate. These include having good support at school or at home, having access to additional specialist resources,

and even in some cases having access to private tutoring. All can increase an individual's chances of success.

PTE Academic introduces a new, additional challenge to the candidate, as it is an entirely online assessment with no human interaction during the test-taking experience. For many candidates, this will be entirely unfamiliar. This is, as the literature on test anxiety reveals, stressful in a range of ways. It has led to the emergence of what is sometimes called a 'shadow' testing industry designed to provide test-specific coaching (Yu & Green, 2021), but also a means of monetising this important aspect of language assessment (Ross and Starling, 2008).

The ways in which students prepare requires careful attention by researchers and test providers because, as research by Symes and Putwain (2016) found, increased test preparation can mitigate not only emotionality, but also physical symptoms such as an increased heart rate, that in themselves might indicate high levels of test anxiety. There is evidence to suggest that students with such high levels of test anxiety spend *more* time preparing for tests than their peers (Culler and Holahan, 1980; Cassady and Johnson, 2002), but while increasing test preparation might reduce physical symptoms, it does not necessarily translate into student attainment. In other words, there is not necessarily a direct correlation between the quantity of personal effort and eventual academic outcome (although there might be in some cases). The situation is significantly more nuanced. Recent research (see O'Sullivan *et al*., 2021) points to a range of other cultural and social differences that relate to preparation and expectations, for example students might take the test purely to 'see what it is like' in the first instance with no expectation of passing. O'Sullivan *et al* also found that some students preferred to practise with a tutor whilst their peers would work alone and feel adequately prepared by simply completing tasks after watching a video. Common to all the students surveyed in their study was preference for focused practice tools and resources to engender a level of preparedness that was appropriate to these settings.

What is potentially emerging from the more recent research in test preparation techniques are two things that relate to an assessment such at PTE Academic.

These are the value of the test in terms of how it shapes future opportunities for the candidate (Richardson, 2022), and the ethics of access to preparation resources. In the latter case, it is because there are cost implications for buying materials, coaching, practice tests etc. (Ross, 2008; Yu and Green, 2021). Given that a potential selling point of more AI-led tests could rest of improved inclusivity for many, the fact that preparation can require the outlay of significant amounts of financial capital is a point worthy of exploration.

**Understanding and beliefs about artificial intelligence**

Public perceptions on the risks and ethics of AI are widely viewed as a potential threat to the public support and validity of AI based systems (Munoz and Maurya, 2022). As large-scale research, such as that commissioned by the European Investment Fund (Atkinson, 2019) has found, public fears are wide-ranging and complex. They comprise generic concerns about the growing societal influence of digital technologies through the datafication of everyday life, the spread of AI, and the use of algorithms (including issues of transparency, accountability, inclusion, bias, ethics, fairness, trust, and privacy).

Specific examples and practices have caught the public imagination and contributed to a negative impression of AI. Some of these relate to assessment specifically, and others to education more widely. One infamous example was the role of algorithms in determining grade allocations for the 2021 cohort of GCSE and A Level candidates in England's summer examination series (2020), the first impacted by the Covid-19 pandemic (H. Richardson, 2020). The awarding process resulted in grade distributions that were coherent at a national level (i.e., they maintained the standard using a comparable outcome model), but they had catastrophic consequences at a local level (M. Richardson, 2020), particularly for candidates from disadvantaged backgrounds. One well publicised example was the award of a 'U' (normally only given if a candidate fails to turn up or fails to complete more or less any answers, so a very unusual event often representing an anomaly). In the centre concerned, a 'U' had been awarded during the previous examination cycle. However, a candidate was selected to be given the grade in 2020 on the basis that because someone had received this grade in a previous year, it needed

to be allocated to the bottom student of the cohort in 2020, even though he/she was not an anomalous candidate and had been predicted to receive a much higher grade.

Another example of an educational AI system with perceived inherent bias is the Intelligent Zoning Engine (IZE), a school place allocation system used since 2017 to determine optimal catchment areas for the Berlin district of Tempelhof-Schoeneberg through calculation of student travel time and distance to school. This has been accused of entrenching disadvantage via inadvertent ghettoization of students from deprived areas of the city (Leaton Gray, 2020). A third example of racial (and sometimes gender) discrimination has been found in biometric facial recognition products, including those sometimes used in schools, which work well in the case of white males, but are frequently less accurate for nearly all other groups (Eubanks, 2018; Raji *et al.*, 2020).

Such negative perceptions and experiences are at odds with what might be viewed as utopian expectations about the potential of AI (Scott, 2017). They also reveal the importance of understanding the fact that no technology can ever be regarded as truly neutral (Leaton Gray, 2020). Even within the AI community, there are significant global concerns about ethics, regulation, and unintended negative consequences, as well and a growing commitment to embedding AI ethics and trust within social policy and regulatory frameworks and professional training. This is reflected in the development of related privacy laws such as the European General Data Protection Regulation [GDPR] (2018) or, in the specific case of children, the United States Children's Online Privacy Protection Act [COPPA] (1998). Public opinion is crucial to the reception and support of AI products because quite apart from any localised ethical considerations, as (Zhang and Dafoe, (2019:187) argue, public trust in institutions, '...can play a major role in shaping the regulation of emerging technologies.' Consequently, there appears to be a strong and growing commitment to promoting the inclusion of ethical commitments and to acknowledge public concerns more transparently within mainstream AI practice. This includes a wide range of deontological, consequential and virtue related ethical concerns (Bartneck *et al.*, 2021) which provide a more rigorous and extended

theoretical and thematic framework than routine validity theory normally considers. Further, it introduces related themes such as concerns about the threats and consequences of automation and the erosion of human agency, risk, trust, transparency, accountability, privacy and psychological consequences that are not usually included in discussions of assessment theory (but should be as assessment embraces AI). Also significant in that literature is ethical attention to the technical and material characteristics of AI i.e., code, sensors, automation, data security (Stahl and Wright, 2018)

Our review identified a rapidly growing literature on AI, trust, and ethics in education (Williamson, 2019; Selwyn and Gallo Cordoba, 2022), and parallel discussions in areas such as health, and interaction with AI based platforms and avatars (Thompson, 2018; Morley et al., 2020; Roski et al., 2021). A distinctive feature of the literature on ethics and trust is the diversity of the themes that it considers. For example, Robinson (2020) focuses on openness and transparency, Qin, Li and Yan (2020) emphasise the importance of functionality and helpfulness of technologies, and individual differences in values and perceptions of technology. Chen et al. (2021) argue that positive user experience and aesthetics impact on public trust, while Aoki (2021) addresses concerns about the erosion of human agency. Aoki (2020) also argues that it is important to communicate with the public about the merits and value of specific AI applications. Dignum (2021:2) explores such ideas further by explaining the societal impact of AI, urging us to appreciate that ’AI systems are more than just the sum of their software components’. They are, she argues, grounded in their specific social contexts and as such, their technical constituents cannot (and should not) be extracted from these settings if we are to interrogate matters of trust.

It is notable that in contrast, discussions of ethics and trust in computer-based testing and AI in education are limited. This suggests that there is a need for more research in these areas because they have potential to raise questions about the extent to which public opinion and media representations of AI inform the perceptions and use of specific AI products, either directly through individual responses, or through impacts on policy and legal frameworks. In the UK, there are

already research institutions (e.g., Department of Digital, Culture, Media and Sport's Office for AI, the Centre for Data Ethics and Innovation, the Ada Lovelace Institute, and the Alan Turing Institute) established to tackle not only the economic potential of AI, but also providing a duty of care about the way that AI impacts on people's lives. This represents a central tenet of this study, given our focus on the lived experience of the candidates.

**Assessment validity and artificial intelligence**

There is a well-established literature and associated professional policy frameworks around validity in educational assessment. In terms of assessment validity, we can make a distinction between the technical and performance related indicators of assessment validity i.e., statistical qualities of tests and items (see Newton, 2007, 2012; Newton and Shaw, 2014), and wider concerns relating to the interpretation, uses and consequences of assessment (e.g., Messick, 1996). Technical literature rarely presents overt discussions of the ethics of testing (admittedly something that is not its focus). However, the importance of validity arguments (Kane, 2015) and consequences (Hubley and Zumbo, 2011) clearly has significance here, because one focus of our PTE Academic candidate research was to consider if the use of AI impacts on any aspects of test validity, depending on the interaction of candidates with the test, *as well* as their overall perception of it. We recognise that perceptions are not a hard indicator of validity issues, rather it is important to acknowledge how a particular understanding of AI might influence candidate views about the efficacy of AI-based examinations. Establishing a clear basis for the use and application of AI in testing designs is valuable, as the misconception is that all computer-based testing systems might use AI is relatively common in educational settings (Leaton Gray and Kucirkova, 2021; Luckin, 2017). if we combine this with fear of the rise of machines (see Richardson & Clesham, 2021) we can see a substantial challenge appearing in relation to just what is being tested, alongside how this is enacted.

Some of the recent literature on computer-based testing and artificial intelligence (Ercikan, 2017; Ercikan, Guo and He, 2020) suggests that digitisation and AI are sources to improve validity, for example through techniques such as the use of real-time process data in adaptive designs, improved construct and score validation

through the use of more granular evidence on response processes and user engagement (see Luckin, 2017). However, there remains a paucity of literature that considers how test users engage with and feel about assessments that make use of artificial intelligence (Zumbo and Hubley, 2017). Given the high stakes nature of the examination, it is evident that there is a transparent need for increased validity to ensure that outcomes are fair and based on ability, as Haladyna and Tindal (2012) suggest. This means that we may not only want to consider the psychometric properties of the test, but also the wider significance of contextual, cultural and consequential factors that shape test reception and performance.

Though previous literature has examined the PTE Academic for construct, criterion and predictive validity (Pae, Greenberg and Morris, 2012; Barkaoui, 2019); few papers have examined the effect of construct-irrelevant variance (CIV) on test outcomes in relation to the PTE Academic. CIV refers to psychological and situational factors that are not intended to be measured during the examination, but nonetheless affect experience, and sometimes even impact the mark or interpretation of an individual's score (Haladyna, Downing and Rodriguez, 2002). To date, research concerning the PTE Academic suggests that there is no difference in exam performance by country of origin or by gender (2012a). This important paper, combined with other works by Pae et al (Moon and Pae, 2011; Pae, Greenberg and Morris, 2012; Pae and Greenberg, 2014) provide a solid foundation in evidencing the PTE Academic as a valid test for the majority of participants. Nevertheless, there is room for further investigation into how diverse groups interact with the PTE Academic tests, as well as to determine what they think. This is important because whilst there is little evidence of CIV, it does not mean that candidates necessarily share this view.

As a computer-based assessment, the PTE Academic has several advantages over paper-based assessments in terms of CIV. The test event is tightly controlled, the measurement of skills is computer assessed and, therefore, relatively impartial, and a broad range of language skills can be assessed (Wise, 2019). However, as argued above, studies on the effects of psychological factors suggest that computer-based assessments do not diminish construct irrelevant factors such as test anxiety

(Hewson and Charlton, 2019). Test anxiety is a form of academic anxiety that is situation specific, and which can have a detrimental impact on behaviours and patterns of beliefs common to testing situations (Cassady, 2004). The impact of test anxiety has been well established in the academic literature. Though a small amount of test anxiety can improve a person's ability to perform, severe forms can significantly affect a person's ability to successfully complete an examination (Kolagari *et al.*, 2018). Perhaps the most establish consequence of test anxiety is an associated decline in performance (Cassady and Johnson, 2002; Putwain and von der Embse, 2018). Students with high levels of test anxiety can experience a decline of cognitive processes, sometimes leading to them failing to achieve the grades they require (von der Embse and Witmer, 2014).

A small body of literature has shown that test anxiety can be specifically linked to language assessments. Aydin's work in Turkey (2009) found that test anxiety prevents language students from reflecting their actual performance in testing situations, reduces studying efficiency and could lead to disengagement from language learning. It is also worth noting that a small body of research has suggested that computer-based language tests, whilst not increasing anxiety overall, can lead to *specific* forms of test anxiety (Jerrim, 2022). A significant contributor to test anxiety in computer-based oral language assessments is the inauthentic modelling of aspects of real life (Butler-Henderson and Crawford, 2020). Research (e.g., (Silfver *et al.*, 2020; Karim Sadeghi, 2022) speculates that this is due to a lack of body language and facial expression to help ease anxiety and encourage learners to express their ideas successfully. Oral communication is, after all, an essentially human act at root, usually relying on multiple stimuli.

The current literature presents inconsistent results in terms of whether computer-based testing causes more test anxiety than traditional paper and pen modalities in Western populations. Several empirical studies (Cassady, Smith and Huber, 2005); Wise, 2019) have found no increase in anxiety when a student takes an exam online, or they have found test anxiety decreases for online assessments. Indeed, a detailed literature review of online examinations (Butler-Henderson and Crawford, 2020) found that all existing literature on student preference indicated that the

majority of students prefer online examinations to paper-based modes. Research by (Woldeab and Brothen, 2019) propose that such inconsistencies occur because for the general population, online assessments are unlikely to increase anxiety. However, a subset of individuals are likely to be adversely affected by online testing and it is important to bear this in mind when researching candidate experiences in a qualitative setting such as this research because it is possible that computer anxiety is a form of CIV that may, on occasion, negatively impact test taking abilities.

The literature in this area of assessment and, in education more broadly, is of course developing apace and so our focus is anchored on candidate perceptions, beliefs and how this aligned with their actual experiences in different test taking settings for PTE. In the next section, we present the ways in which we decided to effectively collect and then analyse data from a range of sources and participants. Our goal was to gain some rich qualitative insights, but from a reasonably sized sample of the PTE test taking population. This meant we had some means of triangulating the anonymous survey responses against in-person interviews to characterise the lived experience of candidates. The research sought to document and reflect upon these authentic assessment experiences and to present an objective view of how AI is viewed, understood and considered in this particular setting.

## The research study

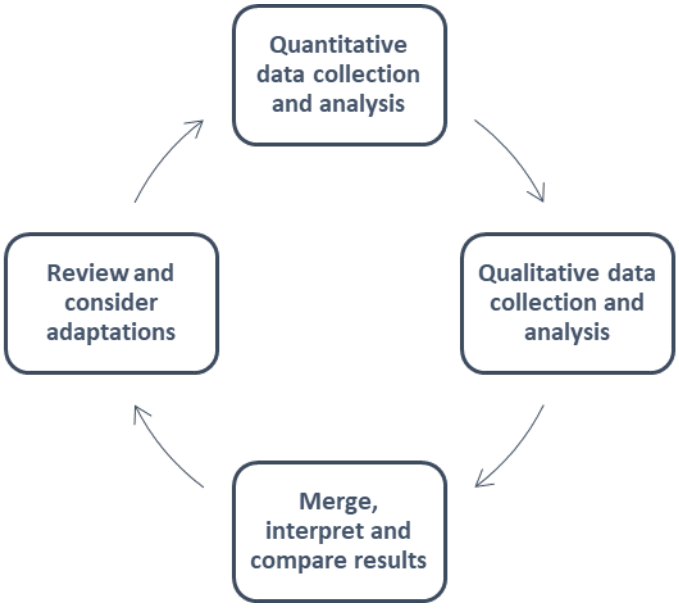
The study's design was guided by the overall research aim to explore and describe the candidates' lived experience with the PTE Academic tests and the three guiding research questions about the role of the candidate, the characteristics of their lived experience and their feedback on the testing cycle. A mixed methods design allowed,

(i) revelation of the broad quantitative patterns in candidates experience with the test, and

(ii) augmentation of insights about the trends with qualitative descriptions of candidates' lived experiences (Creswell, 2018)

This approach allowed us to align categories and themes from qualitative and quantitative data, capturing breadth as well as some depth of data on user interactions. This way of collecting and integrating different data allowed us to build up a comprehensive understanding of users' needs, opinions and experiences relating to the PTE test. The mixed methods design was based on Ivankova, Creswell and Stick's (2006) model of *Sequential Explanatory Design* – see Figure 1 below: a phased approach of collection/analysis of quantitative data followed by a collection/analysis of qualitative data.

*Figure 1: Sequential explanatory mixed-method design over three cycles: October 2021, February 2022, June 2022.*



The former helps the researcher to understand patterns and trends relating to research interests, supplementing them with qualitative data to explain, illuminate and enrich the quantitative. Overall, this approach is meant to provide a deeper understanding of trends and patterns as well as opportunities to explore unexpected results from the quantitative data. The dataset reported here came from two main phases of data collection and analysis. The original research design comprised three cycles of data collection. However, the onset of Covid-19 disrupted access to participants and required a revised ethics application to move data collection entirely online. Given the qualitative nature of the study, we considered the validity of the data collection instruments and whether adaptations[1] were necessary based on the results.

The survey and interviews collected data in two ways, firstly using an online survey tool hosted securely on RedCap (https://www.project-redcap.org - a web application for database and survey data management) within UCL's secure data collection system. The survey, comprising 48 items including attitudinal scales and free text responses, ran from October 2021 to November 2022. The second data collection method was individuals willing to be interviewed for the study. Participants for the survey were sought using a range of approaches that were largely focused online, for example, posting links to the survey on student forums online, using contacts within universities in England to share the link and details of the study, via the Pearson PTE forums and through other social media streams. Potential interviewees were found by asking survey respondents if they would like to talk further (a contact email was provided), through contacts in HEIs and via Pearson PTE[2] databases.

**Ethical Procedures**

The study was approved under the UCL Ethics Committee Review processes in 2021 and assures all participants of anonymity, the right to withdraw data and standard procedures for data storage as defined by the BERA (2018) code of Ethics and UCL data protection regulations. All survey and interview participants signed informed

[1] E.g. Cycle 2 (C2) revealed the need to understand more about how candidates prepared/interacted with the speaking tasks as it had been important in C1 when survey and interview respondents expressed concern about the potential for AI bias based on their accent and intonation. Consequently, the interview guide was adapted to include additional follow-up questions about speaking sections in the test.

[2] Pearson keep databases of PTE candidates who are willing to participate in research so we were able to share the links to the research via these contacts.

consent to participate and received copies of key information about the nature of the study. All participants were assured of their anonymity and confidential handling of data from participation in the survey and/or interviews with researchers.

**Sampling**

The final survey data (see Table 1) resulted in 895 responses; however, the full completion rate was 409; that is, answering every question. Of the remaining 486, there were partial/selective completers (n=77), but another 409 decided not to carry on after reading the project description and consent form. Historically, research suggests this is more common than expected in survey research of this kind (see, Singer, 1978; Sakshaug et al., 2012) who explain that the consent process can be both reassuring and also put prospective respondents off participation. We had to bear in mind the incomplete nature of some responses during the analysis. We add a caveat that there may be gaps in the data; as the results reported as numbers and percentages include those who completed all or most questions (n=486).

*Table 1: Survey responses*

| **Category** | **Count** |
|---|---|
| Answered all questions | 409 |
| Selectively answered questions | 77 |
| Read project description and consent form – did not proceed | 409 |
| Total | 895 |

Most respondents took the test for work and visa purposes (226, 46.8%). The second most common reason for taking the test was to study abroad (148, 30.6%), or to apply for entrance to university (109, 22.6%). In terms of timing, most respondents (408, 88.7%) took the test in November 2021 or later which meant that they had taken the shorter two-hour version of the test. Just 43 (9.3%) respondents took the test in 2019/November 2021 period or before 2019 (9, 1.9%). Given the small numbers who took the longer test, we looked tentatively at whether there were any useful comparative findings between the test takers. These checks revealed nothing substantive in terms of different responses across all of the

themes, so it can be assumed that the analysis and discussion generally focus on candidates who sat a two-hour PTE test.

Almost all respondents (461, 98.3%) took the test at a test centre and just eight (1.7%) took it at home. Most (261, 63%) said it was not the first computer-based test they had taken, but only 149 respondents (30.7%) shared the number of times they attempted the test before they got the score they wanted. Of those who had taken it before, most had two attempts (51, 34.2%). The second largest group took it only once (36, 24.2%) and the remaining respondents took it three (27, 18.1%), four (15, 10.1%), five (8, 5.4%) or more than five times (12, 8.1%). The other language tests tried included IELTS (184, 38.3%), TOEFEL (19, 4.0%), Cambridge ESOL (9,1.9%) and other (27, 5.6%).

21 candidates were interviewed; these included respondents who contacted us after completing the survey as well as others who expressed an interest in talking about their experiences of PTE to Pearson. As Table 2 shows, interviewees initially resided in the countries such as China, Hong Kong, Spain, Nigeria, India, Serbia, UAE, Japan and Brazil, and they had a range of reasons for taking PTE along with varying attempts at taking the test to secure the result they needed.

*Table 2: Summary of interview respondents*

| | | |
|---|---|---|
| Country where test was taken | China | 4 |
| | India | 4 |
| | Australia | 3 |
| | Bosnia & Herzegovina | 1 |
| | Brazil | 1 |
| | Dubai | 1 |
| | Hong Kong | 1 |
| | Japan | 1 |
| | Nigeria | 1 |
| | Saudi Arabia | 1 |
| | Spain | 1 |
| | USA | 1 |
| | Vietnam | 1 |

| | | |
|---|---|---|
| Reason for taking PTE* (some respondents were hoping to work in either UK or Australia, therefore totals exceed 21) | Study Visa (UK)<br>Study Visa (Australia)<br>Study Visa (Norway)<br>Study Visa (NZ)<br>Work Visa (UK)<br>Work Visa (Australia) | 6<br>3<br>1<br>2<br>5<br>6 |
| Times the test was taken | 1<br>2<br>3<br>4<br>6<br>11<br>14 | 12<br>4<br>1<br>1<br>1<br>1<br>1 |

Similarly, to the patterns of survey respondents, their main reasons for taking the test among the interviewees were (i) to study in the UK and Australia or other countries (e.g., New Zealand and Norway) and (ii) work-related visas for Australia, New Zealand and the UK.

Analysis of quantitative data includes descriptive statistics such as frequencies, percentages, and mean values of the respondents' answers, because survey data were predominantly categorical and because we sought rich descriptions of lived experiences without imposing assumptions in the form of quantifiable hypotheses. Alongside the descriptive data was analysis of qualitative data using content analysis of text-based answers. Thematic analysis (Clarke and Braun, 2017) was used to review the interview data and generate recurring themes around five initial overarching categories:

1. Reasons for taking the test
2. Preparation for the test
3. Test-taking experience
4. Test centre experience, and
5. The experience with the AI in the test.

Within each of these categories, we found further distinct themes: Table 3.

*Table 3. Recurring themes in qualitative (interview) data*

| **Category** | **Themes** |
|---|---|
| Reasons for taking the test | Reasons; beliefs and knowledge about the test before |

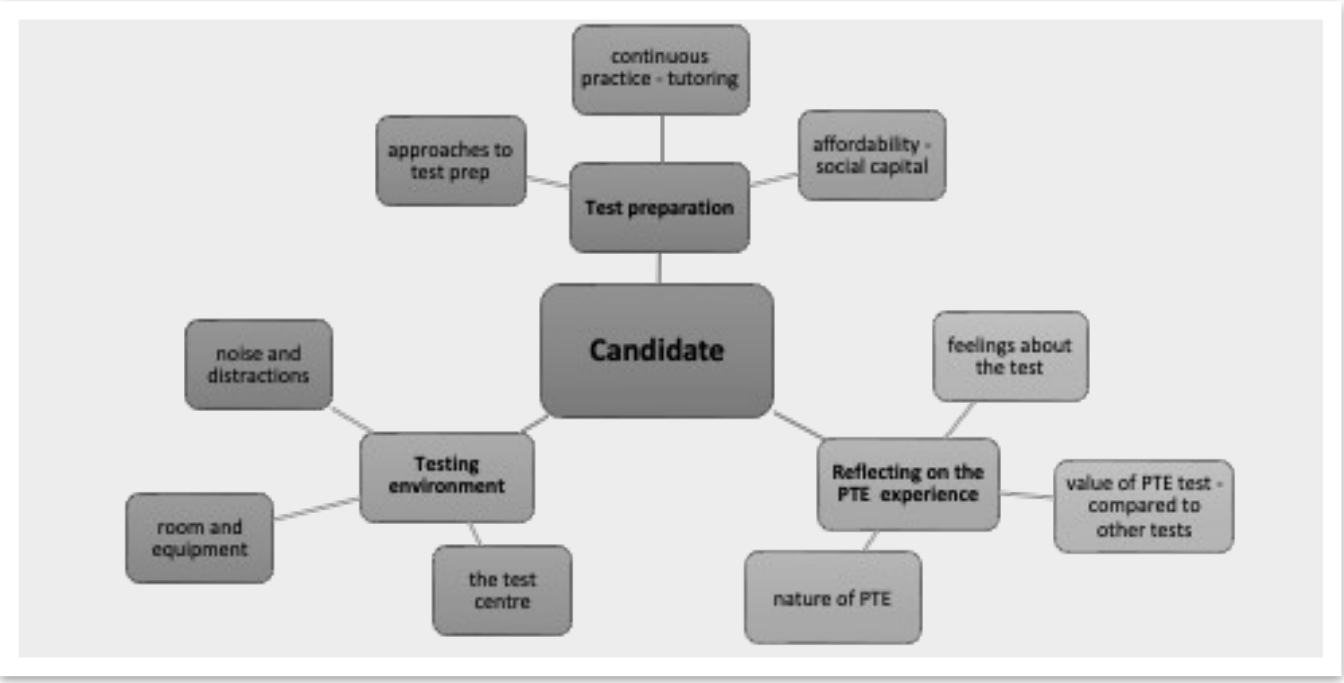
| | |
|---|---|
| | taking the test; comparisons with other tests |
| Preparation for the test | Strategies for test taking; resources and material to prepare; time commitment; advice to other candidates |
| Test-taking experience | Difficulty of questions; difficulty of the testing experience; taking breaks; working with equipment |
| Test centre experience | Check-in process; staff; room; noise; test centre equipment |
| The experience with AI in the test | Strengths and weaknesses; preferences over AI and non-AI English tests; theories about how the AI works; opinions on bias and AI |

The integration of qualitative and quantitative data occurred during the phases of data collection and data analysis and interpretation. In the case of the former, the interviewees were sampled from the pool of survey respondents ensuring that there is an overlap between qualitative and quantitative data. We contacted all the survey respondents within each cycle who expressed an interest in participating in a follow-up interview. In the case of the latter, the questionnaires were analysed through descriptive statistics, content analysis and methods of qualitative categorisation of answers to open-ended questions and the interviewees were analysed qualitatively. Where relevant, in later tables, we provide a low-level quantification of qualitative data such as when counting the frequencies for particular categories of answers in interviews (e.g., the proportion of interviewees who had experience with other tests; the number of times they have taken the PTE Academic and the reasons for taking the test).

## Findings

As the analysis was undertaken, the teams reviewed the research questions to guide organisation and identification of categories; as might be expected from a large and diverse data set the themes were broad, but they were connected and are presented here – Figure 2 – as a summary before more detailed explanations are provided in the subsequent sections of this chapter.

*Figure 2: Categorisation of findings*




The survey and interview results are presented first and followed by the observational data: note that data with % value is from the survey and interviewees are coded a IV1, IV12 etc.

**Resources for preparation**

Given the high-stakes nature of PTE, preparation is key, as we have explained above, and resources to support this are well used. Almost half (214, 49.1%) of the survey respondents said they had bought resources and 222 (50.9%) had accessed the range of free resources online. The websites and platforms respondents used are listed in Table 4 below. The most popular resources for test preparation were language learning resources from private providers, and some respondents (81, 18.7%) used online chat groups and websites to discuss and find out more about the test. The most popular websites for test preparation are YouTube (18, 4%) and www.apeuni.com (13, 2%). The official Pearson website is ranked fifth in the table.

*Table 4: Online resources for test preparation*

| WEBSITE NAME OR URL | Count |
|---|---|
| youtube.com | 18 |
| www.apeuni.com | 13 |
| www.languageacademy.com.au | 7 |

| | |
|---|---|
| https://www.e2language.com/ | 7 |
| **Pearson Official Website** | **6** |
| Telegram Groups and Channels (e.g. AlphaPTE) | 5 |
| WeChat Group Chat | 3 |
| Facebook Groups | 3 |
| https://www.pteacademy.in/ | 2 |
| https://ptetutorials.com/ | 2 |
| www.79score.com | 2 |
| PTE Online Tutorials ; Firefly; Roman PTE Melbourne; Weibo; https://pteplus.com.au/ excellens.com.au; www.pteselfstudy.com; www.fireflyau.com ; https://englishwise.com.au/; PTE Success; www.ptemagic.com.au | 1 |

When asked about their preparation techniques in interviews, respondents said that resources help to set expectations, e.g., what it would be like in the real test situation. Mock tests and templates were deemed a good means of preparation and underline the fact that whilst online testing is becoming widely used in many contexts, within educational settings candidates still have relatively little experience. It is perhaps a factor of this unfamiliarity that led respondents to perceive the real test taking experience as 'harder' and 'different' than the mock tests and they could not always use the speaking templates (IV8). Interviewee 2 said that she had done 'extensive research' to find appropriate advice on taking PTE that was specific to her home region in China to ensure '[I could] find my target region and what they did in the test in March for instance, in Beijing'. She believed this very detailed preparation would make the test taking experience 'predictable' and prepare for questions with 'high frequency' (IV2).

Access to additional support for preparation is of course related to economic factors, and those with more money are able to invest in improving the chances of test success. Among those candidates who paid for resources, the majority invested substantially, with 131 (30.1 %) spending $50-100 and 120 (27.6%) spending >$200. The broad market for preparation for PTE is buoyant as both survey respondents and interviewees said that they had received unsolicited emails or other contact from organisations and/or individuals offering discounted courses and tutorials for PTE as soon as they began to investigate taking the test. This is likely to have been the result of search engine algorithms and internet marketing interacting (Kotras, 2020) to align the candidates with those offering tuition. The resources offered for PTE preparation included private tuition: 140 (41.2%) and practice

resources: 117 (34%). Interviewees 1 and 20 responded to focused email offers and, reported being satisfied with these paid services such as tutorials. Two others (IV 12 and 17) wanted to use Pearson's official resources and purchased mock tests and real sample tests, paying $95 per session for small group tutorials with the PTE Academy and around $140 for additional mock tests.

The most popular preparation tools were the practice resources mentioned above so that the respondents could familiarise themselves with the test and develop test-taking skills specific to PTE Academic. Realistic expectations of the test taking scenario matter a great deal to candidates, and as IV17 claimed, '20% of the success on the test is due to knowing how to pass the test whereas 80% can be attributed to knowledge'. Whilst these quoted percentages don't necessarily reflect the reality of taking a PTE test, such beliefs reflect the long-documented wishes of candidates to control their experience of high stakes tests, see (Harlen, 2008). Sometimes thought of as 'gaming', the candidates seek and practice strategies that they believe will improve their chances of success. For example, IV17 said they had been advised to complete everything in every section of PTE Academic, and that even if they couldn't understand the read aloud sections, it was better to say anything – to 'sing a song or just copy [recite] the lyrics!' Interviewees who had taken other ESL tests were aware of needing different strategies for PTE. One explained that a private tutor had for IELTS, been 'asked to write very long complex sentences in the essay', but when they looked at preparing for PTE, she was told to write shorter phrases and told to make her writing 'more understandable'.

**Practice test resources**

As in most high stakes testing situations, a central preparation strategy is to take as many mock tests as possible. This is because, as prior research demonstrates, practising test taking skills benefits candidates in terms of response speed and can therefore improve self-confidence in the test situation. What is notable about the PTE practice experience is the claims that mock tests were particularly helpful for time management in an online setting. As IV4 said, having only 20 seconds to complete a task demonstrates the value of practising decision making in a short time. However, accuracy is also paramount, and this points to concerns about

whether or not candidates would benefit from being able to touch type. Interviewees 4 and 17 noted this issue saying that they needed to '… type very fast without making spelling mistakes'. IV17 elaborated on the theme of pace saying that the mock tests felt '… relaxed and stretched and the transition from one question to the second one was very nice and smooth', but that the real test situation felt a lot faster and, unsurprisingly, more pressurised.

Interviewees reported that they had practiced taking PTE without timing themselves at all, and once in the test centre, found they were slower than expected, hesitating before speaking and then running out of time. This suggests that perhaps some candidates become overconfident with certain test taking techniques, but that time management is not one of them. This aligns with the research (Stiggins, 1999; O'Sullivan, Dunn and Berry, 2021) that demonstrates the focus on test content above how to spend their time during the exam itself and reinforces the need for candidates to be reminded to prepare in a more holistic way. It would be worth therefore researching the potential differences in practice items and comparable real time items to ascertain differences in time taken to answer. There appears to be little literature that explores these differences in a multi-modal test response environment.

**Question bank resources**

Some respondents accessed question banks that are available through various social media platforms – as shown in Table 4. These 'banks' are compiled by candidates who memorise questions and then share them online. There is an element of generosity in such activity as candidates are trying to '… create a huge data set to help the future students pass the test more smoothly'. Respondents were keen to believe in the efficacy of the content, as IV1 claimed that a PTE question bank she had accessed in China contained items that 'reappeared' in her real test situation, and as IV15 said the evolution of such resources suggests is a need for access to more, free practice content. In terms of learning and providing peer support, we might applaud such behaviours, but what is not clear is the potentially questionable reliability of the test items provided on platforms such as Weibo - https://us.weibo.com/index. They also impact the candidates' expectations

of performance and, as noted in the literature review (Barker, 2020) can increase confidence in ways that are not necessarily demonstrable in the real live test situations.

**Private tutoring**

The use of private tutors to support all phases of education is viewed as a critical part of public-school educational success in most regions of China (Brown and Hirschfield, 2009; Brown and Kong, 2010; Brown, 2018); its prevalence has led to significant policy changes in recent years, notably the Double Reduction policy which has seen the end of tutoring businesses and colleges. This policy does not impact a test such as PTE, and therefore just over 40% of respondents said they had signed up for private classes, coaching sessions and tutorials. Perceptions about the usefulness of these courses are mixed. Valuable advice from private tutors was characterized as learning ways to achieve above their actual ability – an approach to test taking focused on gaining the right score and with little regard paid to their real ability in English. However, the detailed explanations of what happened in private coaching sessions reveal something more than very instrumental approaches to test success. Respondents to the survey and the interviews said they received explicit feedback on their strengths and weaknesses: recommendations made for speaking modules focused on the importance of pace, speed, intonation and pronunciation to ensure good marks. Interviewee 17 claimed that such suggestions 'worked like a charm' as they scored 90 in the speaking module following coaching despite previously having multiple failed attempts.

General comments about commercial language classes involved general guidance about the PTE test as well as material with 'some recommendations on how to answer this question or that question, e.g., how to describe pictures'. However, practising and following the recommendations did not always lead to success on the test, and IV11 said that free YouTube videos with suggested templates for speaking with '[concrete] sentences which you can use to actually describe an image' helped her to develop successful revision and practice strategies. As noted earlier, some candidates took PTE multiple times and not all did this because they had failed. IV11 took it on ten occasions in order to better understand how to increase a

candidate's score. He admitted that he is now a tutor to language candidates himself and uses what he believes to be sage advice about PTE, including keeping talking, and maintaining a consistent speed in your spoken answers. These elements are key, he believed, because the AI might 'hear' differently from a human rater. The same candidate explained how attempts to reverse engineer the PTE test were common in Chinese language schools,

> *Some educational institutions in China, were looking at the test report forms of students, thousands of them... they somehow figure out the algorithm for grading, and what percentage can 'read aloud' contribute to the reading and speaking, and they had a rubric so they visualize that, and that helped them prepare their students.*

Such findings are important not only in terms of what they tell us about test preparation, but also what they say about the beliefs about AI and its role in the PTE test.

**Preparation for the speaking section: feedback apps and templates**

The speaking modules/sections were prioritised for practice with respondents noting that they particularly favoured some mock tests and applications with immediate feedback functionalities to practice for speaking sections on the test. Interviewee 8 used a third-party app for repeating sentences that would provide a score for pronunciation to try and… sound like a native English speaker' the reason being that it was believed she would score better if she sounded 'British'. Such advice left the respondent feeling confused about the efficacy of such advice, and she wondered in seeking speech therapy would improve her chances of success. The evidence (e.g. Tananuraksakul, 2017) about accent does not appear to back up this candidate's beliefs, and instead reveals such seemingly dramatic responses to test taking align with the research evidence, noting the extent to which candidates will go to succeed in such high-stakes settings.

Interviewee 5 reported she avoided using her hands when speaking to enunciate, the advice was 'less hands and more like a robot!'. It was also common to be

advised to adjust voice tone so that candidates were not emphasising particular words in the speaking test. Others (IV20 and IV21) said they practised trying not to pause between words and eliminating so-called fillers such as ‘umm, like’ because they had been advised that this could increase their score. Additional strategies for improving fluency involved using templates for speaking with generic sentences which could be applied to a variety of topics (IV1; IV19).

Candidates typically memorised the models or templates for speaking in the form of off-the-shelf sentences to introduce a topic, e.g. ‘This is a controversial topic still under heated discussion’. Interviewee 19 revealed that after taking the test three times and not being able to get 79+ score in speaking, she was advised and then decided to use a template for speaking. She admitted that she practiced specific ways of explaining things with the aim of ‘feeding’ the computer what it was seeking. She talked about her prior experience of IELTS and said she had used different techniques to try and gain marks, for example, using ‘sophisticated words … thinking ‘oh, my vocab should be really diverse’, whereas her perception of PTE was very different, and she felt that such diversification of vocabulary was less of a priority explaining that she, ‘…was under the impression it's not that important in PTE that fluency kind of comes first.’

Test preparation is of course very important to candidates of PTE and the commitment to attaining high scores and to repeated test taking demonstrated this in our survey and interview respondents. The market for preparatory resources is clearly thriving, and appears to offer a wide range of different types of support. Perhaps quality is an issue too – note the comments regarding the success of crowd created question banks and repetition of items in real tests.

**Taking the test**

In this section we explain the more affective responses to taking the PTE tests; respondents to both the survey and interviews provided a great deal of personal information outlining their feelings, reflections and experiences. The environment for taking any high-stakes test is known to be a potentially important factor in candidate performance, because as research shows (Koretz, 2008), this is critical to

mitigate against any instances of Construct Irrelevant Variance (CIV). Given the reach of PTE, the test centre environments appeared to vary; both survey and interview respondents describing the test centres, employees, the culture/reception at the test centre – all important aspects of understanding their particular experiences.

Most candidates in the survey (367; 88.4%) reported that there was a waiting area at the test centre but almost half (≥48%) said the found the space dark, often cold and crowded and such experiences could increase their anxiety. Most survey respondents (371, 89, 3%), reported that tests started at the expected time and when there were delays, the average waiting time was between 17 and 30 minutes. Another aspect of the test centres were some comments relating to the way candidates were given instructions for taking PTE. Interviewee 5's comment is interesting at it not only appreciates the welcome but suggests an expectation relating to an AI test - '…it's nice to have [the] human touch even though everything is based on AI. It's good to have those live people who welcome you.' Other candidates had mixed experiences; Interviewee 17 (who had two attempts at PTE), compared the experiences:

> *…the first examiner gave us thorough instructions: what are you do, when you [will] finish. The second time, [a] lady with no smile, told us. 'You go there, start; finish, go.*

Interestingly, Interviewee 15 felt very strongly that the online format was 'dehumanising',

> *I felt like I was going to go to prison. They body checked me, checked my documents, I couldn't have my jacket on. I'm not going to see the light of day again!*

Whilst such comments appear to be an effect of the normal anxiety that is a natural part of high-stakes test taking, the data reveal different environmental situations, and these are useful to note in terms of how a candidate might perceive their

experience. The ‘prison’ theme was mentioned by IV4 who felt he was being ‘interrogated’. IV13 described an unusual experience of sitting PTE around one large table and this led to candidates talking loudly.

> *…everyone was very anxious to score more and to be clear they were very loud. Candidates were thinking they couldn't hear themselves because they were wearing masks, they were even louder.*

A noisy test environment is a repeated complaint of PTE candidates; this is characterized by other candidates who are taking the speaking modules at the same time as them. Survey and interview respondents commented on this and the responses are well summarised in IV15’s claim,

> *I do know based on the technology that you have two sets of the microphones, so the AI or the examiner who listens to the recording will be fine because maybe the [other] noise is cancelled. But this affects me. (…) Yeah, I'm trying to read and someone is still speaking… interrupting my focus or reading or reading the materials.*

Interviewee 19 said that the first time she took the test she felt ‘stressed and lost’ because hearing other people answer questions made her confused,

> *‘… in the background a girl answering not the same question, but also something about relatives. Her answer was like ‘sister’ or ‘niece’. I could hear everyone around me, it was distracting and I lost focus a little bit.’*

Most survey respondents (335, 81.8%) had no issues with equipment for the whole of their test experience. When issues were mentions (76, 18.5%), these included furniture (e.g., chair not adjusting) or hardware (e.g., a malfunctioning headset, sticky keyboard). Some respondents reported not being able to hear the sound indicator that signals they can start speaking and others reported some problems with equipment were related to videos – either poor quality, too fast, or hard to understand. At least two interviewees said they found the design of the test to be

‘outdated’ and they noted that during ‘...the Read Aloud sections; it would be helpful to have a countdown timer’. On just two occasions interviewees said that they had said equipment was not working and the exchange of hardware happened before they started their test.

**The physical and emotional experience of the test**

The actual test taking environment was an important factor in this study given the perceived novel experience of taking an online test at an independent centre (that is, not in school or college). Among the survey respondents, the majority said they felt alright (346, 84%) and most (356, 91%) found the chair and table comfortable. Just 66 candidates (16%) reported feeling uncomfortable: and notable issues were increased eye strain (17, 25.4%), and increased feelings of fatigue (13, 18.4%). Anxiety or stress noted by 41 candidates (6.5%), are of course a feature of any high stakes testing situation, but what might be worth deeper exploration is the extent to which a longer screen-focused testing experience potentially exacerbates stress or indeed the physical sensations noted above. The majority of respondents (383, 93.2%) did not take a break during the test (most took the revised two-hour version which, unlike its predecessor has no option for a break) and this is another point for consideration as one survey respondent claimed 'At the centre I went to, they did not allow me to drink water during my exam. I know they were following the process, but it seems insensible'. Usefully, this candidate suggested the option of having a button to pause the test for a self-selecting break; this would of course have implications for the proctoring of the examinations and ensuring security if the candidate leaves the room.

**Perceptions of taking a test with an AI component**

Given the importance of stakeholder endorsement as key element in assessment validation processes (Kane, 2015), there was, amongst the candidates surveyed and interviewed, a strong support for the use of AI in generating test scores. Like any high-stakes test, the way candidates understand interpretation and use of test scores is critical to maintaining trust and it seems that the absence of human scorer bias, the sense of fairness and objectivity in the AI based scoring. However, there were some more specific questions related to the efficacy of that process - the

questions raised by respondents about how well the AI system (and hardware/software) recognised their voice and their accents and these issues are presented below.

The PTE test worked as expected for most respondents (366, 88%); it comprised types of questions they had practised and, positively, almost all survey candidates (390, 94.7%) encountered no surprises in terms of instructions for proceeding through the test and understood each step. Most candidates (319, 77%) were able to complete all test items and finish within the time allowed. When asked about the experience of taking the test on computer, less than half (178, 43.2%) preferred the computer-based tests, but only 10% said they would prefer paper. In terms of perceived difficulty, one third of respondents (139, 33.7%) said they found the test difficult, and harder than they expected, with only 8% saying that the test was easier than they expected (33, 8.0%). However, about a third of respondents also found that taking the test was less stressful than they had expected (129, 31.3%). Some of the words used to characterize their experiences included ‘intense’, ‘very tiring’ and IVV4 was very specific claiming that PTE requires ‘concentrating on the screen’ that requires a lot of ‘working memory and a ‘different level of stress’. When compared to other tests, I4 believed that the experience of taking PTE was less intense than GRE language tests, but more intense than IELTS – others said it was as fast and intense as TOEFL.

**Perceptions of the fairness of the test**

We were interested in whether candidates believed that PTE allows them to demonstrate their competence and ability in English and (333, 81.2%) of survey respondents felt that it did. Whilst this data and the fact 344 survey respondents (82.95%) believed PTE is a fair test, it is important to consider how this aspect of the test-taking experience is intertwined with candidates’ beliefs about bias and fairness and the use of AI. During the interviews, we were able to delve a bit deeper into candidate perceptions and this revealed some challenges to the perceived fairness of PTE. We know that such qualitative data is not generalisable, but it does provide a richer insight to personal experience and highlights misconceptions about the use of AI and how some PTE test takers might understand

its value and limitations. For example, one interviewee (IV20) felt confused during the speaking tasks and wondered ‘Did the system listen to me properly? Has my voice reached them or not?’ Given the general trend to trust an AI rater compared to a human being (Thompson, 2018; Roski et al., 2021), this interviewee (and others who were interviewed) appeared to find the lack of a response from an automated system as detached and remote. This view was also echoed by others with claims such as that AI marking is unreliable because of the way that AI evaluates a speech by recognising keywords without looking for ‘coherent sentences’ (IV1) and without knowing and understanding contextual meaning.

Perhaps unsurprisingly, given its high-stakes, the temptation to explore potential gaming of the PTE test featured in the data about fairness. Some of the free text responses from the survey provide examples of how candidates believe they might gain advantages notably learning ‘tricks’ such as keeping talking to gain more marks (even if the sentences are unintelligible), or memorising templates from preparation sites to recall in situ, or memorising answers to the online question banks, as mentioned earlier. One interviewee (IV1) believed that the AI features for PTE have ‘blind spots’ around which they could create answers to enhance their score – this particular candidate believed that their PTE score was higher than their actual ability. We are unable, within the constraints of this kind of research, to substantiate such claims, but they present further ideas about how AI and the candidate are supposedly interacting.

Two other important themes relating to fairness could be of concern and these relate to the quality of sound recordings (can candidates be sure that in all test centres, all audio equipment functions to the same level?), and perhaps most importantly, the additional noise of other candidates in the testing centres. The interviewees were able to explain how they felt it important to manage the environment carefully, for example IV17 reported that some candidates waited for others to stop speaking before they recorded their responses to the read aloud sections. This candidate also found the noise of others completing read aloud impacted on their concentration in the subsequent writing sections '… [when others start] speaking, you really lose focus’. As Saenger (1982) reminds us, being in a

room with others reading or speaking aloud independently and simultaneously is something that has not been common since the days of medieval monasteries (with the exception of learning taking place within specially designed and soundproofed language laboratories, which these rooms are usually not). These days, the usual practice is to learn languages silently, unless the learner is part of a deliberate conversation with others in the room or repeating something together with others after the teacher. Therefore, it is unsurprising that PTE candidates find noise intrusion a challenge, given that this is a pedagogical practice not normally associated with modern learning and assessment practices. This problem may be exaggerated if a candidate suffers from any kind of auditory processing problem or hearing impairment (Pham and Karuza, 2022) therefore we suggest that it is worth reviewing considering access and arrangements for candidates.

Only a small number (27, 6.6%) of survey respondents had requested access arrangements: these included time adaptation (12, 44.4%), adapted equipment (11, 40.7%), screen adaptation (9, 33.3%) and an Interpreter (5, 18.5%). Of these 27 candidates, ten (37%) had their request for the access support refused and this left them feeling frustrated about the overall test experience. One interviewee claimed that her request for additional time for reading was not allowed and despite being in a separate room away from other candidates, she was interrupted by '… all the people outside talking'. This is a single instance but led to the candidate choosing to leave the PTE test and take IELTS instead because they provided a better time accommodation. These snapshots are valuable in terms of thinking about the candidate experience overall and how different candidates can be best supported; these data are unrelated to the AI experience and the next section moves from the actual reflections on being in the test situation and on to beliefs about the technology.

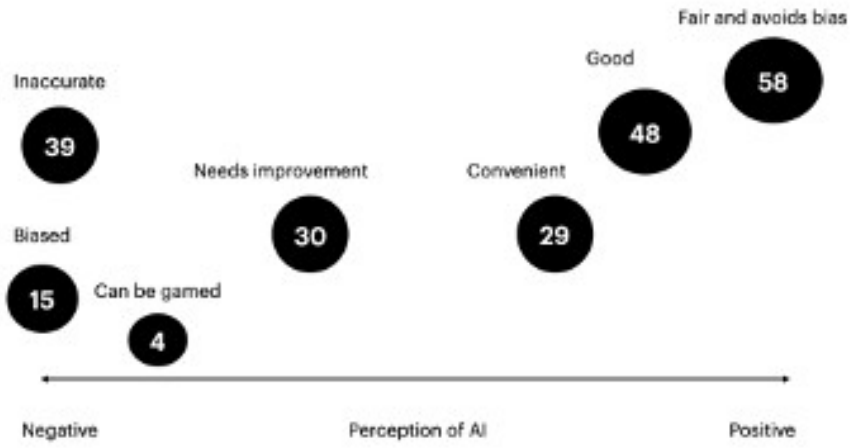
**Perception of the AI for automated scoring**

When asked about their actual results we found that 239 (59.3%) of all survey respondents (n=406) got the results they needed, but more interestingly most (209, 52.8%) received the results they expected or higher. In what follows there are discussions about how students viewed their scores and specifically the impact that

some believed the AI components of PTE had on their results. and about half (187, 47.2%) had results that were lower than they expected.  What is important to note here is that whilst a score might be lower than expected, it does not mean the candidate did not achieve what they needed.  The literature on candidate expectations reveals common patterns globally where test takers are keen to score only the highest marks in all their tests (Richardson, 2022); such behaviours of course lead to a skewed sense of success and are likely to impact their overall views of how much they value a test experience.

The range of views on AI and its use in the context of taking a PTE test are summarised in Figure 3 below revealing a broadly positive perception with some fears of impartiality and inaccuracy.

*Figure 3: Summary of survey responses to the use of AI in the PTE tests*




We asked survey respondents if they knew that PTE had some AI functions and the majority, 328 (81%) knew that the test employs AI to create the final score; just one person claimed that they were expecting a human examiner and were '... surprised to know an AI assessed my test!'. They were more excited by the prospect of something novel and mentioned the AI in ways that reflect some of the literature (Pelau, Dabija and Ene, 2021) now exploring how people view AI and other

technology that might assume human behaviours. The characterisation of the use of AI in assessing PTE was enthusiastic. Forty-seven (12%) of survey respondents explained their support for the use of AI in testing by describing it variously as ‘ground-breaking’, ‘innovative’ and ‘a technological advancement’. Respondents explained that they felt the non-human functions for marking would mean the test was less biased and fairer; as one interviewee explained, working on screen means ‘… avoiding human examiners [is a good thing], especially for those who are stressed, introverted and nervous’. A further 58 (14%) claimed that the convenience of AI was valuable, particularly the speed and efficiency of taking the test and receiving the test results.

What is striking in this part of the study was the extent to which candidates believed that the AI content of PTE had more of an influence that it does in terms of the score they received. For example, IV17 said that ‘…even if you can’t hear what the computer is asking just say something, the AI will give you some points’ and believed that it was possible to game this section of the test because the AI functions would be listening to them in a way similar to a human listener. Other negative perceptions mentioned in the survey included beliefs that AI technology is fundamentally flawed (30, 7.3%), that it was unable to assess genuine English language skills (30, 7.3%) and that it will be biased and discriminatory (15 3.6%). Respondents who characterized AI as a weak factor in the PTE tests highlighted a range of concerns including beliefs that an automated scoring system might inaccurately assesses language skills - specifically the assessment of speaking. Survey respondents described the automated scoring as ‘immature’ and ‘incorrect’ and a few (15, 3.6%) believed that the AI may be biased by “privileging certain accents, intonations and voices at the expense of others”. One interviewee claimed that they had done some personal research (unpublished), and this had led them to believe it was simply not possible for a computer to reliably measure spoken language competency. They argued that competency in speaking in a computer-based test relied on volume making, leading to their conclusion that PTE is unfair. Whilst personal opinions might be eschewed in favour of evidence-based studies, the critical point here is the ways in which understanding (or misunderstanding) of

AI capabilities will, for some candidates, influence trust in the experience of taking a computer-based test with AI components.

Nevertheless, what might be termed more radical views are difficult to determine because as research shows us (see, (Kampmeyer, Matthes and Herzig, 2015; Barker, 2020) it's common for candidates to over predict their expected grades and believe they have done better than they have. As one interviewee claimed, 'I don't believe my assessment results were correct although I passed the exam comfortably. I was seeing low scores in sections where I am the most confident and sure of my capabilities and performance in the exam.' Such claims can be an effect of overinflated self-efficacy, and they reflect the contemporary issues in how candidates characterise themselves as successes or failures (Richardson, 2022) within test taking scenarios. The way in which candidates conceptualise their success is not necessarily a part of the PTE preparation processes or resources, but what quotes such as the one above suggests is that despite passing (and being able to move to the next goal, e.g., university), there is scepticism in the value and even validity of the scores. Such beliefs about the test are also intertwined with the high-stakes nature of its outcome and to the commercial nature of this kind of testing. Interviewee 19 raised this issue and likened Pearson to a kind of testing version of Google and was critical of the cost (in her case, $400) related to language testing for university entrance and citizenship/employment.

**Candidates' comparisons of PTE Academic and IELTS**

Some survey respondents claimed that PTE avoids human error '...as seen with IELTS' so we explored in more depth the perceived differences between the PTE and IELTS tests in interviews with those interviewees who had had the experience of both tests. Firstly, we found a theme of convenience; for example, Interviewees 12 and 15 mentioned the speed with which the results are released as key reasons for choosing it instead of IELTS. The second theme related to preparation; Interviewees 6 and 16 said that they found PTE easier than IELTS in terms of preparation; IV6 took IELTS to study in Australia, but then used PTE for a work visa because as she explained, '[IELTS] takes longer to prepare, especially in my view it is writing and listening that takes a lot more time to prepare, if you want to achieve the

equivalent score’.  Similarly, Interviewee 10 said the time it took to prepare for the test was one of the main reasons for choosing PTE and this respondent cited the computer-based nature of the test as important,

> *‘... of course, when you speak to a real person, the way they assess it’s different. For PTE It’s more about being precise and for IELTS you’re going to have to be more conscious about which is a better way of delivering words.’*

Interviewee 1 believed that PTE was more ‘predictable and controlled’ with fewer new and unpredictable questions meaning that they felt less tired by the test experience of PTE compared to IELTS. A third theme in comparing PTE and IELTS was the feedback on outcomes and how the results are reported; respondents appreciated the detailed breakdown from PTE tests. As Interviewee 12, said, knowing the details of vocabulary and grammar (in speaking) helped them to create an accurate, “...assessment of what's the strength and what's the nature of your knowledge. And it's helpful if I want to get a better score for the future if I want to improve my English”. The perceptions of PTE are a crucial factor in how we address the questions guiding this research and as will be outlined in the final sections of the report, present a range of useful experiential reflections that suggest ways to improve how well the PTE tests are understood as well as the expectations of those taking them.

## Conclusions

This research was guided by three central questions and, as the preceding sections have demonstrated, through a mixed methods approach, we have been able to provide a detailed characterisation of the candidates' experiences, perceptions and views of PTE.

### The role of the candidate in the AI-led language testing setting of PTE Academic

By role here we mean in the sense of preparing for, and taking, the test. The studies conducted by UCL were able to interrogate some of the detailed preparation that candidates typically undertake for PTE. The survey data set of over 400 complete returns offers evidence and insights into the actions and beliefs of a range of test takers. In addition, the detailed interviews with candidates add a further layer of insights to inform our characterisation of taken the PTE tests. Central to the data are some key findings relating to access and fairness in relation to test preparation and whilst they continue to echo the research that focuses on test preparation broadly, they also point to different ways of seeing test preparation when it's (a) in an online setting, and (b) when the candidate has misunderstood the role of the technology, in this instance, the application of AI and how it interacts with their role as a test taker.

Throughout all data collection phases, the respondents noted practical issues that had impacted the quality of their experiences, and these influenced their beliefs about the test. Some of these issues included equipment, but others related to individual needs of candidates such as time allowances and so on that are now a standard expectation for a candidate with a specific learning difficulty, for example. Given the structure and design of PTE, there are of course limitations to how the test taking experience might be adapted for candidates, but this remains a prominent issue for Pearson as it is necessarily linked to trust in the test and how it is perceived.

Another issue that we raise here relates to human capital and social advantage embedded within the entire process of taking PTE. The respondents highlighted

potential discrimination for those from socio-economically disadvantaged backgrounds, given the amount of money commonly spent on buying support for practice, tuition and so on. This demonstrates the persistent link between access to financial capital and test preparedness. Given the high-stakes nature of PTE, it is unsurprising that candidates will do as much as they can to ensure success, and the use of skilled tutors speaks to the ways in which students feel they need preparation and who/what they are willing to trust in such matters. However, the more problematic side of this finding is the extent to which other potential candidates might be disadvantaged in their preparation if they are unable to access funds to (a) pay for more bespoke support and (b) pay for multiple attempts – especially as some of them feel that PTE is a more challenging test than many of its competitors and that they need to take practice iterations to get a 'real' feel for the experience.

The actual role of AI in the PTE tests is clearly misunderstood by many candidates and whilst they claim to know what the test is and does, they do not always appear to grasp the precise role played by AI during the process. In terms of our original research project title, The Future of the Artificially Intelligent Examination, this lack of knowledge about the capacity and/or extent of AI is important. During the writing up of these data, a new AI tool, ChatGPT, appeared for trial on the internet. It is capable of writing almost instantaneous answers to questions – and of course, learning while responding from a global dataset of users. The speed and effectiveness of such technology is held in awe and even some fear, so deciding how to create a reassuring message for users relating to AI is important and, going forward, something that might need continuous reflection, adaptation, and opportunities for discussion with candidates. The rhetoric surrounding the use and application of tools such as ChatGPT (Lund and Wang, 2023) reminds us that there is work to be done in terms of helping create public discourses that seek opportunities as well as being wary of how they might negatively impact educational settings.

**Characterisations of the lived experience of candidates in AI-led language test domains**

The most striking findings in relation to this question are the candidates' beliefs about the extent to which the AI components of PTE were interacting with them during the testing experience. The evidence from this study demonstrates there is interest in and support for the use of AI technologies in some forms of testing and that it is perceived as being fairer, or at least as fair as human judges. What is of most interest to us are those candidates who felt a sense of disconnect from their experience (particularly the survey respondents), those who felt that they were being watched, that the entire test centre experience was 'cold' and as one put it, 'inhuman'. Whilst there is no doubt that taking a high stakes test is likely to invoke feelings of high stress, it appears the perception of how they are/are not interacting with technology is something that could benefit from better exploration. The nature of how AI devices appear to those interacting with them is part of a growing literature and it is worth considering the extent to which candidates consider the PTE tests to have anthropomorphic characteristics that really 'listen' to them or are affected by their accent/way of speaking (Pelau, Dabija and Ene, 2021). Knowing more about this aspect of the candidates' beliefs could be applicable to the preparation resources and the actual test centre experience; more broadly it will add to the literature on assessment in testing situations.

The study reflect ways that candidates interact with the online nature of the tests and reveal assumptions about the nature of AI in testing, and particularly how AI is used in the PTE tests. Many of the assumptions were incorrect and this demonstrates gaps between the actual experience and perceived AI capabilities within the specific context of assessment. The context of *how* AI is used appears to be a potentially useful avenue for further exploration because it might be that candidates trust the AI to complete the final scoring processes correctly, whereas they are less trusting of its accuracy in determining their ability from a recording. What these findings suggest is that more research is necessary to explore the candidates' knowledge about how assessment technology works, what its limits and capabilities are, and when it does or does not interact with a candidate during the test experience. For many years, the research has suggested that assessment has

to move beyond simply moving a paper test behind a glass screen, but there is also further work to be done in explaining the way that certain types of assessment work: how they are collecting evidence for particular skills, e.g. speaking, and most importantly how they are doing this comparably to a live situation with a human examiner. Such foci all relate to the perceived validity of the tests – the overall results gleaned by candidates from their visit to the test centre; given that validity is at the heart of assessment process and practice, how it is conceptualised and understood by candidates in this new setting is important.

**Documenting candidates' beliefs**

We would like to emphasise the importance of including the candidates in all development of PTE going forward, not only to improve and keep updating the test design, but also to ascertain what they know and think about new technological developments in online testing, and in the ways that AI is used in those test experiences. Much of the misinformation and incorrect assumptions could be corrected and explored with regular input from candidates and by asking them specific questions relating to their experiences. The details of being at a test centre and how that feels are important here too. This was heightened by the misunderstanding of instructions, the intense focus necessary in a screen-based test and the timing/clock that is ever present.

Candidates were grateful to have a chance to talk about their experiences and wanted to explain issues they had encountered as well as offering suggestions and praise. It would seem there is fine line between taking a high stakes test that is innovative, accessible, provides results promptly and appears contemporary, and then anxieties about perceived fairness, concerns that the technology could be faulty (interestingly few said that this could apply to human beings) and the impersonal nature of a test that involves speaking aloud and listening to recorded voices.

An important theme for the research team from the outset was thinking beyond the capturing of a candidate's day at the test centre, and considering the nature of the assessment process and how this might be better represented as feedback for the

PTE development teams. The following points are an initial attempt to summarise some recommendations for further discussion and exploration in this final report on the two-year study.

## Recommendations

a. There is a need to document and challenge the commercial preparation industry, especially regarding its effectiveness and how truthfully it presents itself to candidates, as well as assessing how far it align to candidates' requirements and wishes.

b. There could be investigation into the impact of socio-economic disadvantage on test preparation opportunities with the aim of supporting democratic access to the test and providing good quality preparatory materials to all candidates.

c. There should be further research into extending the exploration of candidates' understanding of the role and potential for AI in assessment.

d. There could be further investigation into how AI capabilities within the PTE tests are used and the extent to which candidates know how they are interacting with AI during and after their testing experience.

e. Given the pace of change in the use of AI tools and the public interest in new technology such as ChatGPT, our findings suggest a need for more public discussion about the limits and opportunities of using AI in educational assessment beyond a test such as PTE. Pearson could consider holding an event to explore myths associated with the use of AI in assessment, and to explain the realities and benefits of its use.

## References


American Psychological Association, author *et al.* (eds) (2014) *Standards for educational and psychological testing / American Educational Research Association, American Psychological Association, National Council on Measurement in Education.*

Aoki, N. (2020) 'An experimental study of public trust in AI chatbots in the public sector', *Government information quarterly*, 37(4), p. 101490. Available at: https://doi.org/10.1016/j.giq.2020.101490.

Aoki, N. (2021) 'The importance of the assurance that "humans are still in the decision loop" for public trust in artificial intelligence: Evidence from an online experiment', *Computers in human behavior*, 114, p. 106572. Available at: https://doi.org/10.1016/j.chb.2020.106572.

Atkinson, R. (2019) *Don't fear AI (Volume 2)*. European Investment Bank.

Aydin, K.B. (2009) 'Automatic thoughts as predictors of Turkish university students' state anxiety', *Social behavior and personality*, 37(8), pp. 1065–1072. Available at: https://doi.org/10.2224/sbp.2009.37.8.1065.

Barkaoui, K. (2019) 'Examining sources of variability in repeaters' L2 writing scores: The case of the PTE Academic writing section', *Language testing*, 36(1), pp. 3–25. Available at: https://doi.org/10.1177/0265532217750692.

Barker, I. (2020) 'Could exam results hinge on confidence?', *The Times Educational Supplement*.
Bartneck, C. *et al.* (2021) 'An Introduction to Ethics in Robotics and AI', *Springer Briefs on Ethics*, pp. 1–114. Available at: https://doi.org/10.1007/978-3-030-51110-4.

Bennett, R.E. (2015) 'The Changing Nature of Educational Assessment', *Review of Research in Education*, 39, pp. 370–407. Available at: https://doi.org/10.3102/0091732X14554179.

Brown, G. (2018) *Assessment of student achievement*. (Ed psych insights).
Brown, G.T.L. and Hirschfield, G.H.F. (2009) 'Students' conceptions of assessment: Links to outcomes', *Assessment in Education: Principles, Policy & Practice*, 15(1), pp. 3–17. Available at: https://doi.org/http://dx.doi.org/10.1080/09695940701876003.

Brown, G.T.L. and Kong, H. (2010) 'The Validity of Examination Essays in Higher Education: Issues and Responses equ_460 276..291'. Available at: https://doi.org/10.1111/j.1468-2273.2010.00460.

Butler-Henderson, K. and Crawford, J. (2020) 'A systematic review of online examinations: A pedagogical innovation for scalable authentication and integrity', *Computers and education*, 159, pp. 104024–104024. Available at: https://doi.org/10.1016/j.compedu.2020.104024

.
Cassady, J.C. (2004) ‘The influence of cognitive test anxiety across the learning–testing cycle’, *Learning and instruction*, 14(6), pp. 569–592. Available at: https://doi.org/10.1016/j.learninstruc.2004.09.002.

Cassady, J.C. and Johnson, R.E. (2002) ‘Cognitive Test Anxiety and Academic Performance’, *Contemporary educational psychology*, 27(2), pp. 270–295. Available at: https://doi.org/10.1006/ceps.2001.1094.

Cassady, J.C., Smith, L.L. and Huber, L.K. (2005) ‘Enhancing validity in phonological awareness assessment through computer-supported testing’, *Practical assessment, research & evaluation*, 10(18), p. 18.

Chassignol, M. *et al.* (2018) ‘Artificial Intelligence trends in education: A narrative overview’, in *Procedia Computer Science*. Available at: https://doi.org/10.1016/j.procs.2018.08.233.

Chen, T. *et al.* (2021) ‘AI-based self-service technology in public service delivery: User experience and influencing factors’, *Government information quarterly*, 38(4). Available at: https://doi.org/10.1016/j.giq.2020.101520.

Clarke, V. and Braun, V. (2017) ‘Thematic analysis’, *The journal of positive psychology*, 12(3), pp. 297–298. Available at: https://doi.org/10.1080/17439760.2016.1262613.

Creswell, J.W. (2018) *Research design : qualitative, quantitative, and mixed methods approaches*. Fifth edit. Edited by J.D. Creswell. Thousand Oaks, California: SAGE Publications, Inc.

Culler, R.E. and Holahan, C.J. (1980) ‘Test anxiety and academic performance: The effects of study-related behaviors’, *Journal of educational psychology*, 72(1), pp. 16–20. Available at: https://doi.org/10.1037/0022-0663.72.1.16.

Dignum, V. (2021) ‘The Myth of Complete AI-Fairness’, in *Artificial Intelligence in Medicine*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 3–8. Available at: https://doi.org/10.1007/978-3-030-77211-6_1.

von der Embse, N.P. and Witmer, S.E. (2014) ‘High-Stakes Accountability: Student Anxiety and Large-Scale Testing’, *Journal of applied school psychology*, 30(2), pp. 132–156. Available at: https://doi.org/10.1080/15377903.2014.888529.

Ercikan, K. (2017) *Validation of score meaning for the next generation of assessments : the use of response processes*. Edited by Kadriye. Ercikan and J.W. Pellegrino. New York: Taylor & Francis (The NCME Applications of Educational Measurement and Assessment Book Series). Available at: https://doi.org/10.4324/9781315708591.

Ercikan, K., Guo, H. and He, Q. (2020) ‘Use of Response Process Data to Inform Group Comparisons and Fairness Research’, *Educational assessment*, 25(3), pp. 179–197. Available at: https://doi.org/10.1080/10627197.2020.1804353.

Eubanks, V. (2018) *Automating inequality : how high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press.

Haladyna, T.M., Downing, S.M. and Rodriguez, M.C. (2002) 'A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment', *Applied Measurement in Education* [Preprint]. Available at: https://doi.org/10.1207/S15324818AME1503_5.

Haladyna, T.M. and Tindal, G. (2012) *Large-scale Assessment Programs for All Students*. Taylor and Francis. Available at: https://doi.org/10.4324/9781410605115.

Harlen, W. (2008) *Student assessment and testing / edited by Wynne Harlen*. Edited by W. Harlen. Los Angeles ; London: SAGE (Sage library of educational thought and practice).

Hewson, C. and Charlton, J.P. (2019) 'An investigation of the validity of course-based online assessment methods: The role of computer-related attitudes and assessment mode preferences', *Journal of computer assisted learning*, 35(1), pp. 51–60. Available at: https://doi.org/10.1111/jcal.12310.

Hubley, A.M. and Zumbo, B.D. (2011) 'Validity and the Consequences of Test Interpretation and Use', *Social indicators research*, 103(2), pp. 219–230. Available at: https://doi.org/10.1007/s11205-011-9843-4.

Ivankova, N. V., Creswell, J.W. and Stick, S.L. (2006) 'Using Mixed-Methods Sequential Explanatory Design: From Theory to Practice', *Field methods*, 18(1), pp. 3–20. Available at: https://doi.org/10.1177/1525822X05282260.

Jerrim, J. (2022) 'Test anxiety: Is it associated with performance in high-stakes examinations?', *Oxford review of education*, ahead-of-print(ahead-of-print), pp. 1–21. Available at: https://doi.org/10.1080/03054985.2022.2079616.

Kampmeyer, D., Matthes, J. and Herzig, S. (2015) 'Lucky guess or knowledge: a cross-sectional study using the Bland and Altman analysis to compare confidence-based testing of pharmacological knowledge in 3rd and 5th year medical students', *Advances in health sciences education : theory and practice*, 20(2), pp. 431–440. Available at: https://doi.org/10.1007/s10459-014-9537-1.

Kane, M.T. (2015) 'Explicating validity', *Assessment in Education: Principles, Policy & Practice*, 23(January), pp. 1–14. Available at: https://doi.org/10.1080/0969594x.2015.1060192.

Karim Sadeghi (2022) *Technology-Assisted Language Assessment in Diverse Contexts*. Taylor and Francis (Routledge Research in Language Education). Available at: https://doi.org/10.4324/9781003221463

Kolagari, S. *et al.* (2018) ‘The Effect of Computer-based Tests on Nursing StudentsÂ’ Test Anxiety: a Quasi-experimental Study’, *Acta informatica medica*, 26(2), p. 115. Available at: https://doi.org/10.5455/aim.2018.26.115-118.

Koretz, D.M. (2008) *Measuring up : what educational testing really tells us*. Harvard University Press.

Kotras, B. (2020) ‘Mass personalization: Predictive marketing algorithms and the reshaping of consumer knowledge’, *Big data & society*, 7(2), p. 205395172095158. Available at: https://doi.org/10.1177/2053951720951581.

Leaton Gray, S. (2020) ‘Artificial intelligence in schools: Towards a democratic future’, *London Review of Education*, 18(2), pp. 163–177. Available at: https://doi.org/10.14324/LRE.18.2.02.

Leaton Gray, S. and Kucirkova, N. (2021) ‘Ai and the human in education: Editorial’, *London Review of Education*, 19(1), p. Editorial. Available at: https://doi.org/10.14324/LRE.19.1.10.

Luckin, R. (2017) ‘Towards artificial intelligence-based assessment systems’, *Nature Human Behaviour* [Preprint]. Available at: https://doi.org/10.1038/s41562-016-0028.

Lund, B.D. and Wang, T. (2023) ‘Chatting about ChatGPT: how may AI and GPT impact academia and libraries?’, *Library Hi Tech News*, ahead-of-p(ahead-of-print). Available at: https://doi.org/10.1108/LHTN-01-2023-0009.

Madaus, G.F. (1988) ‘The Distortion of Teaching and Testing: High-Stakes Testing and Instruction’, *Peabody Journal of Education*, 65(3), pp. 29–46. Available at: https://doi.org/10.1080/01619568809538611.

Messick, S. (1996) *Validity and washback in language testing.* Edited by Educational Testing Service and Graduate Record Examinations Board. Princeton, N.J.: Educational Testing Service (Research report (Educational Testing Service) ; RR-96-17).

Moon, Y. and Pae, J.-K. (2011) ‘Short-term Effects of Automated Writing Feedback and Users’ Evaluation of Criterion’, *Applied Linguistics*, 27(4), pp. 125–150. Available at: https://www.dbpia.co.kr/Journal/articleDetail?nodeId=NODE01848264 (Accessed: 5 September 2020).

Morley, J. *et al.* (2020) ‘The ethics of AI in health care: A mapping review’, *Social science & medicine (1982)*, 260, pp. 113172–113172. Available at: https://doi.org/10.1016/j.socscimed.2020.113172.

Munoz, J.M. and Maurya, A. (2022) *International perspectives on artificial intelligence*. 1st ed. Edited by J.M. Munoz and A. Maurya. London, England: Anthem Press.

Newton, P.E. (2007) ‘Clarifying the purposes of educational assessment’, *Assessment in Education: Principles, Policy & Practice*, 14(2), pp. 149–170. Available at: https://doi.org/10.1080/09695940701478321.

Newton, P.E. (2012) ‘Questioning the Consensus Definition of Validity’, *Measurement*, 10(1–2), pp. 110–122. Available at: https://doi.org/10.1080/15366367.2012.688456.

Newton, P.E. and Shaw (2014) *Validity in educational & psychological assessment*. Edited by S.D. Shaw, associated with work Cambridge Assessment, and University of Cambridge. Local Examinations Syndicate. London: Sage Publications.

O’Sullivan, B., Dunn, K. and Berry, V. (2021) ‘Test preparation: an international comparison of test takers’ preferences’, *Assessment in Education: Principles, Policy and Practice*, 28(1), pp. 13–36. Available at: https://doi.org/10.1080/0969594X.2019.1637820.

Pae, H.K. and Greenberg, D. (2014) ‘The Relationship Between Receptive and Expressive Subskills of Academic L2 Proficiency in Nonnative Speakers of English: A Multigroup Approach’, *Reading psychology*, 35(3), pp. 221–259. Available at: https://doi.org/10.1080/02702711.2012.684425.

Pae, H.K., Greenberg, D. and Morris, R.D. (2012) ‘Construct Validity and Measurement Invariance of the Peabody Picture Vocabulary Test-III Form A’, *Language assessment quarterly*, 9(2), pp. 152–171. Available at: https://doi.org/10.1080/15434303.2011.613504.

Pelau, C., Dabija, D.-C. and Ene, I. (2021) ‘What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry’, *Computers in human behavior*, 122, p. 106855. Available at: https://doi.org/10.1016/j.chb.2021.106855.

Pham, C.T. and Karuza, E.A. (2022) ‘Noise-induced differences in the complexity of spoken language’, *Quarterly journal of experimental psychology (2006)*, pp. 174–248. Available at: https://doi.org/10.1177/17470218221124869.

Putwain, D.W. and von der Embse, N.P. (2018) ‘Teachers use of fear appeals and timing reminders prior to high-stakes examinations: pressure from above, below, and within’, *Social Psychology of Education*, 21(5), pp. 1001–1019. Available at: https://doi.org/10.1007/s11218-018-9448-8.

Raji, I.D. *et al.* (2020) ‘Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing’, in *AIES, New York*, pp. 145–151. Available at: https://doi.org/10.48550/arxiv.2001.00964.

Razavipour, K., Habibollahi, P. and Vahdat, S. (2021) ‘Preparing for the higher education admission test: preparation practices and test takers’ achievement goal orientations’, *Assessment and evaluation in higher education*, 46(2), pp. 312–325. Available at: https://doi.org/10.1080/02602938.2020.1773392.

Richardson, H. (2020) *Exam results: Where did it go wrong and what happens next? - BBC News*, *BBC News*. Available at: https://www.bbc.co.uk/news/education-53811391 (Accessed: 5 September 2020).

Richardson, M. (2020) *A-level debacle has shattered trust in educational assessment*, *The Conversation*. Available at: https://theconversation.com/a-level-debacle-has-shattered-trust-in-educational-assessment-144640.

Richardson, M. (2022) *Rebuilding public confidence in educational assessment*. London: UCL Press.

Richardson, M. and Clesham, R. (2021) 'Rise of the machines? The evolving role of AI technologies in high-stakes assessment', *London review of education*, 19(1). Available at: https://doi.org/10.14324/LRE.19.1.09.

Robinson, S.C. (2020) 'Trust, transparency, and openness: How inclusion of cultural values shapes Nordic national public policy strategies for artificial intelligence (AI)', *Technology in society*, 63. Available at: https://doi.org/10.1016/j.techsoc.2020.101421.

Robson, C. (2002) *Real World Research: A Resource for Social Scientists and Practioner-Researchers*, *Blackwell Publishing*. Available at: https://doi.org/10.1016/j.jclinepi.2010.08.001.

Roski, J. *et al.* (2021) 'Enhancing trust in AI through industry self-governance', *Journal of the American Medical Informatics Association : JAMIA*, 28(7), pp. 1582–1590. Available at: https://doi.org/10.1093/jamia/ocab065.

Ross, J.A. and Starling, M. (2008) 'Self-assessment in a technology-supported environment: the case of grade 9 geography', *Assessment in education : principles, policy & practice*, 15(2), pp. 183–199. Available at: https://doi.org/10.1080/09695940802164218.

Ross, S.J. (2008) 'Language testing in Asia: Evolution, innovation, and policy challenges', *Language testing*, 25(1), pp. 5–13. Available at: https://doi.org/10.1177/0265532207083741.

Saenger, P. (1982) 'Silent reading: Its impact on late medieval script and society', *Viator*, 13, pp. 367–414.

Sakshaug, J.W. *et al.* (2012) 'Linking Survey and Administrative Records', *Sociological methods & research*, 41(4), pp. 535–569. Available at: https://doi.org/10.1177/0049124112460381.

Scott, D. (2017) *Education Systems and Learners*. London: Palgrave Macmillan UK.
Selwyn, N. and Gallo Cordoba, B. (2022) 'Australian public understandings of artificial intelligence', *AI & society*, 37(4), pp. 1645–1662. Available at: https://doi.org/10.1007/s00146-021-01268-z.

Silfver, E. *et al.* (2020) ‘Classroom bodies: affect, body language, and discourse when schoolchildren encounter national tests in mathematics’, *Gender and education*, 32(5), pp. 682–696. Available at: https://doi.org/10.1080/09540253.2018.1473557.

Singer, E. (1978) ‘Informed Consent: Consequences for Response Rate and Response Quality in Social Surveys’, *American sociological review*, 43(2), pp. 144–162. Available at: https://doi.org/10.2307/2094696.

Stahl, B.C. and Wright, D. (2018) ‘Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation’, *IEEE security & privacy*, 16(3), pp. 26–33. Available at: https://doi.org/10.1109/MSP.2018.2701164.

Stiggins, R. (1999) ‘Assessment, student confidence and school success’, *Phi Delta Kappan* [Preprint]. Available at: https://doi.org/http://dx.doi.org/10.1108/17506200710779521.

Symes, W. and Putwain, D.W. (2016) ‘The role of attainment value, academic self-efficacy, and message frame in the appraisal of value-promoting messages’, *British journal of educational psychology*, 86(3), pp. 446–460. Available at: https://doi.org/10.1111/bjep.12117.

Tananuraksakul, N. (2017) ‘Building up Thai EFL students’ positive attitudes toward their non-native English accented speech with the use of phonetics website’, *Teaching English with technology*, 17(4), pp. 52–63.

Thompson, S. (2018) ‘Modelling Trust Between Users and AI’, in *Artificial Intelligence XXXV*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 171–176. Available at: https://doi.org/10.1007/978-3-030-04191-5_14.

Williamson, B. (2019) ‘Policy networks, performance metrics and platform markets: Charting the expanding data infrastructure of higher education’, *British journal of educational technology*, 50(6), pp. 2794–2809. Available at: https://doi.org/10.1111/bjet.12849.

Wise, C.N. (2019) *Assessment and Instruction for Developing Second Graders’ Skill in Ascertaining Word Meanings from Context*.

Woldeab, D. and Brothen, T. (2019) ‘21st Century Assessment: Online Proctoring, Test Anxiety, and Student Performance’, *International Journal of E-Learning & Distance Education*, 34(1), pp. 1–10.

Yu, G. and Green, A. (2021) ‘Preparing for admissions tests in English’, *Assessment in Education: Principles, Policy and Practice*, 28(1), pp. 1–12. Available at: https://doi.org/10.1080/0969594X.2021.1880120.

Zhang, B. and Dafoe, A. (2019) ‘U.S. Public Opinion on the Governance of Artificial Intelligence’, in *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics*, pp. 187–193. Available at: https://doi.org/10.48550/arxiv.1912.12835.

Zumbo, B.D. and Hubley, A.M. (2017) *Understanding and investigating response processes in validation research*. Edited by B.D. Zumbo and A.M. Hubley. Cham, Switzerland: Springer Nature (Social indicators research series, volume 69).