Research Note: Establishing Construct and Concurrent Validity of Pearson Test of English Academic

Ying Zheng
Pearson, London, UK
Ying.zheng@pearson.com

John H.A.L. De Jong
Pearson, London, UK
John.dejong@pearson.com

March 2011

1. Introduction

Pearson Test of English Academic (PTE Academic) is a computer-based international English language test launched globally in 2009. The purpose of the test is to assess English language competence in the context of academic programs of study where English is the language of instruction. This paper reports the processes involved in collecting evidence to support the validity claims of the test for that purpose. The paper begins with a review of the definitions of and the threats to construct validity; defines the construct validity of PTE Academic; and then reports both qualitative and quantitative evidence (including research methods, results, and revisions) which Pearson have gathered to support the construct validity of PTE Academic. In addition, the concurrent validity of PTE Academic is presented by comparing PTE Academic with other externally established criteria or tests.

2. Defining Construct and Concurrent Validity

Achieving test validity is an essential concern in test development, particularly when a test is used for high-stakes purposes. However, as Messick commented 'many test makers acknowledge a responsibility for providing general validity evidence of the instrumental value of the test, but very few actually do it' (Messick, 1992, p. 18). More recently, Weir (2005) reported that while most examinations claim different aspects of validity, they often lack validation studies of actual tests that demonstrate evidence to support inferences from test scores.

Messick's (1995) unified view of validity predicated that validity is a multifaceted concept, which can only be established by integrating considerations of content, criteria, and consequences into a comprehensive framework for empirically testing rational hypotheses about score meaning and utility. It is widely recognized that the validation process should start from the very beginning of test development. Schilling (2004) maintained that, in addition to a posteriori validity evidence (which traditionally focused on scoring validity, criterion-related validity and consequential validity); a priori validity evidence (such as test design decisions and the evidence that supports these decisions) also makes a significant contribution to the establishment of validity. Similarly, Weir (2005) highlights the importance of a priori validity evidence when he stated that 'the more fully we are able to describe the construct we are attempting to measure at the a priori stage, the more meaningful might be the statistical procedures contributing to construct validation that can subsequently be applied to the results of the test' (p. 18), because the statistical analysis at a posteriori stage do not generate conceptual labels by themselves, and therefore to make the scores meaningful, the test developers can never escape from the need to define what is being measured at the beginning of test development.

2.1 Construct validity

The process of establishing construct validity for any test should be an on-going endeavour in which 'various sources of evidence are gathered, synthesized, and summarized' (Cizek, Rosenberg, & Koons, 2008, p. 298) from the very beginning of the test development process so as to arrive at an integrated evaluatory judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores (Messick, 1989). In the same paper, Messick (1989) identified two main threats to construct validity:

- Construct under-representation: This occurs when the test fails to capture important aspects of the construct that it is intended to measure or when part of the construct is not present in the test.
- Construct-irrelevant variance: This occurs when the test scores are influenced by factors irrelevant to the construct. For example, an individual's background knowledge, personality, characteristics, test-taking strategies, and general intellectual or cognitive ability must all be construct-irrelevant to this test and effort needs to be made to keep influences such as these to a minimum.

To avoid construct under-representation, a test should establish a strong link between a test score and what it claims to measure. There should be a wide range of realistic tasks to ensure sufficient coverage of target language use (TLU) situations, with the correspondence to TLU situations being closely related to the notion of authenticity. Bachman (1991) explains that there are two types of authenticity: situational authenticity and interactional authenticity. Situational authenticity refers to the extent to which the test tasks simulate the characteristics of TLU tasks. Interactional authenticity refers to the extent to which knowledge, skills, and cognitive processes critical in the TLU situation also represent those in the construct definition and those required to perform well in the situations about which test performance is intended to generalize.

To avoid construct-irrelevant variance, test items should be scrutinized in order to check whether certain groups of test takers are advantaged or disadvantaged by the test as a result of their cultural and educational background. The target test-taking group of this test is heterogeneous. Test takers speak a wide variety of first languages, have a wide variety of cultural and social backgrounds, and come from, or intend to study in, a wide variety of academic disciplines. If, therefore, a task requires topical knowledge which is not shared by all test takers, then bias is present. The solution is to ensure that, for all items, the probability of providing the correct answer is dependant only on knowledge that everyone shares, regardless of their geographical, ethnic and cultural origins, or on information that has been provided in the stimulus and is thereby available to all candidates.

2.2 Concurrent validity

Concurrent validity is the degree to which results from a test agree with the results from other measures of the same or similar constructs. The problem, however, with this type of validity evidence, as Moller (1982) reminds us, is that we need to check whether or not the test or criterion is valid. If they are not valid or not designed to measure the same construct, then, one cannot claim that a test has criterion-related validity because it correlates highly with another test or external criterion of performance.

In the next sections, supporting evidence for PTE Academic construct validity as well as concurrent validity are presented.

3. Supporting Evidence for PTE Academic Construct Validity

PTE Academic measures English language proficiency for communication in tertiary level academic settings. It is targeted at intermediate to advanced English language learners, which means that the test needs to assess critical writing, listening and reading skills such as the understanding of subtle aspects of texts and implied meanings. In order to claim that PTE Academic is fit for its purpose, a variety of validity evidence has been collected from the various stages of test development through to its administration. The constructs measured are the communicative language skills needed for reception, production and interaction in both oral and written modes, as these skills are necessary to successfully follow courses and to actively participate in the targeted tertiary level education environment.

This section reports on the development steps taken and on how, at each stage, evidence was collected to support the validity claims of the test. The first stage covers the development of the most fundamental document, the test specification, which provides a clear blueprint to item writers, markers and test users on what PTE Academic is and how it assesses academic capability. The second stage concerns item writing training. Item writers were among the earliest groups of people who were provided with the test specification and the training program aimed to familiarise item writers with both the PTE Academic test specification and the CEF (Common European Framework of Reference for Languages). The third stage initiated an item peer review, an extra expert scrutiny process, and a three-stage sensitivity review, to validate the domain representation. The construct validation procedures also included empirical validation of data from two field tests (stage 4 and stage 5). Field Test One provided qualitative data to inform the test development team about a variety of issues, such as test format, instructions, time limit of certain item types in the test and computer devices. After making adaptations to these issues, Field Test Two was administered. Its statistical analysis involved two parts. First, native speaker performance was compared to non-native speaker performance in terms of required response time and response correctness. Secondly, item-level analyses were performed to analyse field test data using Rasch/Partial Credit modelling. Fit indices were used to evaluate the item qualities. The following sections outline in detail the supporting evidence for PTE Academic construct validity.

3.1 Developing test specifications

Since the test scores will be used for university admission purposes, the high-stakes nature of the decisions require this test to be valid for the inferences the test users make, that is, whether test takers have adequate English proficiency to succeed in English-medium tertiary settings. In developing valid test items, quality assurance measures were adopted at each stage of the test development processes on the basis of test specification, which serves as an operational definition of the constructs intended to measure. An outline of the test specification of PTE Academic is presented below.

Table 1: Test specification for PTE Academic

Purpose of the instrument
The test will measure English language proficiency for communication in English-medium tertiary setting.
Construct or domain that will be measured
Communicative language skills will be assessed for reception, production and interaction in the oral and written modes as these skills are needed to successfully follow courses and actively participate in education and training where English is the language of instruction.
Framework of the instrument
• Presentation mode: oral-audio or video, graphic, or any combination thereof; • Content of stimulus: such as instruction, definition, explanation, description, argument • Task type: such as retrieving information, interpreting, deducting, combining information from different sources, evaluating • Required competences: such as grammatical, lexical, phonological, pragmatic, strategic • Response mode: oral or written • Response format: such as multiple choice (single or multiple answers), clicking hotspots in texts, selecting from drop down lists, drag and drop, highlight, short answers, extended response
Text length
The test will take up to 3 hours.
Context in which the instrument is to be used
English-medium education and training and related professional fields
Characteristics of intended participants
Learners of English as a second or other language who are applying for admission to courses where English is the language of instruction or admission to professional bodies or skilled professions
Psychometric properties
The items will all have a demonstrable and robust relationship with the reporting scales towards which they purport to contribute. Scores on reporting scales will have a standard error of measurement no greater than 15 percent of the score range covering 68% of the intended target population
Conditions and procedure of administering the instrument
Tests will be administered on computer in dedicated Pearson test centers
Procedures of scoring
Depending on the item format, some items will be scored dichotomously, but the majority of items will be scored polytomously. Item scores for closed response will be generated by machine, whereas item scores for constructed, open responses will be generated by automatic scoring system trained on initial human ratings.
Reporting of the results
Overall, communicative skills and enabling skills scores will be reported as a profile of the candidate's level of ability. In addition scores will be provided in numerical form and in relation to the CEF.

3.2 Developing item types

To mitigate test format effects, as well as to emulate the range of functions and situations that students may encounter when pursuing academic studies in English, 21 item types were designed at the beginning stage, with one item type being excluded during the data analysis procedure. The 20 item types, testing different skills or combinations of skills, were presented in different formats, modes of delivery, response modes, and task types. Tables 2 to 5 show the item types for the different sections of PTE Academic. The table describes each task and the traits which are scored and lists how each task contributes to PTE Academic scores. In addition to overall score, a score profile is developed, including four communicative skill scores and six enabling skill scores. The four communicative skills are listening, speaking, reading, and writing. The six enabling skills are grammar, oral fluency, pronunciation, spelling, vocabulary, and written discourse.

Table 2: PTE Academic Speaking Section

Item type	Task	Traits scored	PTE Academic scores contributed to	
Read aloud	A text appears on screen. Read the text aloud	Content, Oral Fluency, Pronunciation	Overall Score	
			Communicative Skills scores:	Reading Speaking
			Enabling Skills scores:	Pronunciation Oral Fluency
Repeat sentence	After listening to a sentence, repeat the sentence	Content, Oral Fluency, Pronunciation	Overall Score	
			Communicative Skills scores:	Listening Speaking
			Enabling Skills scores:	Pronunciation Oral Fluency
Describe image	An image appears on screen. Describe the image in detail	Content, Oral Fluency, Pronunciation	Overall Score	
			Communicative Skills scores:	Reading Speaking
			Enabling Skills scores:	Pronunciation Oral Fluency
Re-tell lecture	After listening to or watching a lecture, retell the lecture in your own words	Content, Oral Fluency, Pronunciation	Overall Score	
			Communicative Skills scores:	Reading Speaking
			Enabling Skills scores:	Pronunciation Oral Fluency
Answer short question	After listening to a question, answer with a single word or a few words	Scored either correct or incorrect depending on appropriateness and accuracy	Overall Score	
			Communicative Skills scores:	Listening Speaking
			Enabling Skill scores:	Vocabulary

Table 3: PTE Academic Writing Section

Item type	Task	Traits scored	PTE Academic scores contributed to	
Summarize written text	After reading a passage, write a one-sentence summary of the passage	Content, Form, Grammar, Vocabulary	Overall Score	
			Communicative Skills scores:	Reading Writing
			Enabling Skills scores:	Grammar, Vocabulary
Write essay	Write an essay of 200-300 words on a given topic	Content; Development, Structure and Coherence; Form, General linguistic range, Grammar usage and mechanics, Spelling, Vocabulary range	Overall Score	
			Communicative Skill scores:	Writing
			Enabling Skills scores:	Grammar, Vocabulary, Spelling Written Discourse

Table 4: PTE Academic Reading Section

Item type	Task	Scoring method	PTE Academic scores contributed to	
Multiple- choice, choose single answer	After reading a text, answer a multiple-choice question on the content or tone of the text by selecting one response	Response scored either correct or incorrect	Overall Score	
			Communicative Skill score:	Reading
Multiple- choice, choose multiple answers	After reading a text, answer a multiple-choice question on the content or tone of the text by selecting more than one response	Partial credit given for each correct response	Overall Score	
			Communicative Skill score:	Reading
Re-order paragraphs	Several text boxes appear on screen in random order. Put the text boxes in the correct order	Partial credit given for any correctly ordered pair	Overall Score	
			Communicative Skill score:	Reading
Reading: Fill in the blanks	A text appears on screen with several blanks. Drag word or phrases from the blue box to fill in the blanks	Partial credit given for selecting each correct response for a blank	Overall Score	
			Communicative Skill score:	Reading
Reading and writing: Fill in the blanks	A text appears on screen with several blanks. Fill in the blanks by selecting words from several drop down lists of response options	Partial credit given for selecting each correct response for a blank	Overall Score	
			Communicative Skill scores	Reading Writing

Table 5: PTE Academic Listening Section

Item type	Task	Traits scored	PTE Academic scores contributed to	
Summarize spoken text	After listening a recording, write a summary of 50-70 words	Content, Form, Grammar, Vocabulary, Spelling	Overall Score	
			Communicative Skills scores:	Listening Writing
			Enabling Skills scores:	Grammar, Vocabulary Spelling
Multiple-choice, choose multiple answers	After listening to a recording, answer a multiple-choice question on the content or tone of the recording by selecting more than one response	Partial credit given based on the relative order of each adjacent pair of sentences	Overall Score	
			Communicative Skill score:	Listening
Fill in the blanks	The transcription of a recording appears on screen with several blanks. While listening to the recording, type the missing words into the blanks	Each gap is scored based on the correctness of each word	Overall Score	
			Communicative Skills scores:	Listening Writing
Highlight correct summary	After listening to a recording, select the paragraph that best summarizes the recording	Response scored either correct or incorrect	Overall Score	
			Communicative Skills scores:	Listening Reading
Multiple-choice, choose single answer	After listening to a recording, answer a multiple-choice question on the content or tone of the recording be selecting one response	Response scored either correct or incorrect	Overall Score	
			Communicative Skills scores	Listening
Select missing word	After listening to a recording, select the missing word or group of words that completes the recording	Response scored either correct or incorrect	Overall Score	
			Communicative Skill score:	Listening
Highlight incorrect words	The transcription of a recording appears on screen. While listening to the recording, identify the words in the transcription that differ from what is said	Partial credit given for each correct word from the audio transcript	Overall Score Communicative Skills scores:	Listening Reading
Write from dictation	After listening to a recording of a sentence, type the sentence	Partial credit given for each correct response	Overall Score	

3.3 Item Writer Training

Based on the test specification, items were developed, piloted and analyzed. Item writers, who were commissioned to write test items, were the focus of the second stage of item development. This section reports on item writer training including: item writer recruitment; developing Item Writer Guidelines; checklists for item writers; writing to CEF levels; and procedures for monitoring the performance of item writers.

3.3.1 Item Writer Recruitment

PTE Academic assesses international English, which is defined as English that is readily understandable by other speakers of English. Item development teams were established in the USA, the UK, and Australia and sourced test material from these countries and from other international contexts.

To ensure the quality of test items, item writers must meet specific minimum standards of qualifications and experience. They must

- have native fluency in English (with at least 10 years of education in an English-speaking country)
- have very good verbal communication skills (including writing, reviewing and editing skills)
- have an undergraduate or higher degree in applied linguistics, English language and literature, education, or a closely related field
- be computer literate

Other recommended qualifications and experience include:

- Knowledge of testing programs, policies, and standards such as the CEF
- Experience as an item, passage or educational material writer for ESL or EFL tests
- Experience in teaching ESL or EFL

3.3.2 Item Writer Training

Qualified item writers were trained to become familiar with the Item Writer Guidelines, which included the detailed test specification of PTE Academic and the CEF scales. As mentioned previously, the test specification is one of the most important reference documents that item writers should consider when they start work as it defines the constructs which the test intends to assess and makes the scores of PTE Academic 'meaningful' to its users.
Item Writer Guidelines were developed based on the initial specification document. The Guidelines specified in much greater detail the characteristics of each item and gave item writers rules and checklists to ensure that a high proportion of their items were fit for purpose and suitable for inclusion in the item bank.

Tables 6 to 8 illustrate the main structure of the Item Writer Guidelines for each of the four skills. To develop reading and listening items, item writers were largely trained in three aspects: 1) target language use situation; 2) selecting appropriate reading or listening texts; and 3) the CEF scale on reading and listening.

'Target Language Use Situation' aims to inform writers what skills or abilities PTE Academic intends to assess in reading and listening, and by what method (i.e., the format of the item types). The second part of the Guidelines explains the characteristics of reading and listening passages through which test takers can best demonstrate their abilities. For the reading items, this includes test sources, authenticity, discourse type, topic, domain, text length and cultural suitability. For the listening items, it includes text sources, authenticity, discourse type, domain,

topic, text length, accent, text speed, how often the material will be played, text difficulty, and cultural suitability.

Since both speaking and writing items elicit productive skills, the Guidelines explain target language use situation with details of the CEF scale from levels B1 to C2. In the Guidelines for writing, the purpose of writing discourse and the cognitive process of academic writing are presented in a matrix format with recommendations for preferred item types. The purposes of writing tasks are defined as 1) to reproduce, 2) to organize or reorganize, and 3) to invent or generate ideas. Three types of cognitive processing are differentiated: to learn; to inform; and to convince or persuade. In the Guidelines for speaking, item writers are instructed to produce topics focusing on academic interests and university student life. A list of primary speaking abilities is also provided, including the ability to comprehend information and deliver such information orally, and the ability to interact with ease in different situations.

Table 6: Structure of PTE Academic Item Writer Guidelines

Reading		
Target Language Use Situation: What abilities/ skills/ knowledge are necessary and how do they need to be assessed to establish the construct validity argument for PTE Academic?	**List of abilities PTE Academic aims to assess in Reading Tasks**	1. Identify the main ideas and supporting details 2. Understand the author's purpose, technique, attitude and rhetorical intent 3. Precisely understand details including facts, reasons, outcomes, hypothesis, evidence implications 4. Infer the meaning of unfamiliar lexical items 5. Understand conceptual themes and concepts (e.g., cause-effect, compare-contrast, cause-result) 6. Understand synaptic structure, discourse makers, lexical and/ or grammatical cohesion 7. Extract salient details to summarize 8. Reading critically in an in-depth appreciating style, draw logical inferences, evaluate and challenge hypothesis and evidence 9. Integrate information from multiple sources into a coherent whole, and generate an organizing frame that is not explicitly stated
	Range of techniques employed in this test to assess the abilities mentioned above: Multiple-choice, Multiple-response, Clicking on hotspots, Ordering sentences, Gapped texts, Summary, Matching paragraphs (integrating listening and reading)	
Selecting appropriate Reading Text: What kinds of reading materials provide test takers with the best chance to demonstrate their abilities?	**Text sources**	Texts of Academic Interests/ Tests related to All aspects of student life
	Authenticity	Should be excerpted from real-life texts. Do not simplify or adapt the text
	Discourse type	Descriptive and Instructive Reading / Informational and Expository Reading / Persuasive and Argumentative Reading
	Content Domain	Educational/ Academic
	Topic	Topics cover subjects in arts, science, social science/humanities and business administration
	Text length	Please refer to the item specifications for the desired passage length for each item type
	Difficulty of texts	Reader variables/ Text variables/ Tasks variables
	Cultural suitability	Cultural-neutral
CEF Scale on Reading from Levels B1 to C2		

Listening		
Target Language Use Situation: What abilities/skills/knowledge are necessary and how do they need to be assessed to establish the construct validity argument for PTE Academic?	**List of abilities PTE Academic aims to assess in Listening Tasks**	1. Identify the purpose and scope of lecture/speeches 2. Understand the main ideas and supporting ideas, gist, implications 3. Detect the tone and attitude of the speaker 4. Identify text structure and connection between parts 5. Understand the communicative function of utterances, and techniques that the speaker uses to convey the message 6. Infer the conceptual framework and relationships within discourses (e.g., generalization, conclusion, cause-effect) 7. Extract salient points to summarize the oral discourses 8. Draw valid references and conclusions about the speaker's intent or the general context 9. Listen critically in an in-depth appreciating style, draw logical inferences, evaluate and challenge hypothesis and evidence 10. Integrate information from multiple sources into a coherent whole, and generate an organizing frame that is not explicitly stated
Range of techniques employed in this test to assess the abilities mentioned above: Different response formats are used. The test taker may be directed to respond orally, in writing, or by reading options. The response type may be selected, as in a multiple-choice task; limited, as when a single word or short phrase is required; or extended, as in a summary.		
Selecting appropriate Listening Text: What kinds of listening tasks provide test takers with the best chance to demonstrate their abilities?	**Text sources**	Texts of Academic Interest / Texts Related to Student Life
	Authenticity	Should be genuine and contain oral features. Scripts texts are usually coherent and polished, not recommended
	Discourse type	Descriptive and Instructive / Informational and Expository / Persuasive and Argumentative
	Domain	Educational/ Academic
	Topic	Topics cover subjects in arts, science, social science/humanities and business administration
	Text length	Please refer to the item specifications for the desired passage length for each item type
	Accent/ standard	Accents can be categorized as standard or regional English (American, British and Australian)
	Text speed	Normal or fast
	How often played	Each recording is played only once
	Format	Audio/Video: Aural text can be accompanied by visuals which provide information about the setting and

		give a sense of where the language is taking place. Context visual usually facilitate comprehension
	Text difficulty	Text characteristics/ Task characteristics
	Cultural suitability	Cultural-neutral

CEF Scale on Listening from Levels B1 to C2

Writing	**Models of Writing discourse**			
Target language Use Situation: At postgraduate level, students need to produce a range of text types integral to academic performance	Domain intention/ purpose / Cognitive processing	Reproduce	Organize/reorganize	Invent/ Generate
	To learn	Listening & Writing: Fill in the blanks Writing: Write from dictation		
	To inform (inferential)		Reading & Writing: Fill in the blanks Listening & Writing: Summarize spoken text	
	To convince or persuade			Reading & Writing: Summarize written text Writing: Write essay
What Makes a Task Difficult?	Variables related to the task itself, such as topic, the expected discourse mode of the response, the variable related to the scoring processes such as the background and experience of the raters, the nature of the rating scale, and training			

CEF Scale on Writing from Levels B1 to C2

Speaking		
Target language Use Situation	**Topics**	**Academic Interests:** Students need to speak to achieve academic purposes, such as participating in class discussions, responding to professors' questions, or giving presentations **University student life:** Events that occur on campus (e.g. bookstore, cafeteria, housing office, library, medical services)
	List of abilities	**Language competency**: pronunciation and intonation **Content:** the ability to comprehend information, and deliver the information orally **Fluency:** the naturalness of speech production, the degree to which comprehension is impeded by hesitancy, distraction and inappropriate silence **Strategic capacity:** achievement strategies and restructuring **Oral interaction:** the ability to interact with ease in different situations
CEF Scale on Speaking from Levels B1 to C2		

3.3.3 Checklists for Item Writers

When item writers are familiarized with the Guidelines, they are not only informed 'what construct the items should assess and in what format', but are also instructed on how to prevent their items from being labeled as 'construct-irrelevant' and/or 'construct under-presented'. The most relevant Guidelines are presented below:

To write Reading items, item writers are trained to

- Sample as many different texts and topics as possible. Allowing a wide range of topics to be covered reduces the potential bias from a restricted range of topic areas.
- Choose texts from general readings rather than specialized textbooks. The texts should be relatively non-technical and able to be understood by a general audience.
- Ensure that chosen texts chosen are of an appropriate level of difficulty, as estimated by the CEF scale.
- Be aware that item discrimination in a reading item describes the ability of the item to distinguish good readers from poor readers. An item that discriminates well between students of different ability levels is a good item.
- Ensure that the information required to solve the task is stated in or can be implied from the text. Test takers have to read and understand the relevant paragraphs and should not be able to get the item correct from world knowledge alone.
- Use more global questions which ask test takers to synthesize information or draw conclusions at a subtle level. Avoid questions which ask for a superficial understanding of clearly stated information.
- Focus on important information in the text, rather than trivial information.

To write Listening items, item writers are trained to

- Ensure good acoustics and minimal background noise so that test takers can hear clearly and comfortably.
- Avoid highly decontextualized and truncated texts as these can be very different from what happens in a target language use situation.
- Be aware that, in many cases, visual information serves to increase the cognitive load of the test taker. Visual information must, therefore, be straightforward and easy to process.
- Avoid testing aspects of listening comprehension that are open to alternative interpretations.
- Ensure that test takers have the necessary background knowledge to finish each task. Use texts that are dependent on knowledge that everyone has regardless of their geographical, ethnic and cultural origins, or on information that has been provided in the stimulus.

To write Speaking items, item writers are trained to

- Ensure that the sample produced by test takers can be scored according to the rating criteria.
- Make sure that it is possible to make inferences from the scores to the construct.
- Avoid selecting poor quality drawings or photographs as this affects the difficulty level of tasks.

To write Writing items, item writers are trained to

- Be aware that it is important that all tasks are constructed carefully to allow test takers to perform to the best of their abilities. Eliminate variation in scoring that can be attributed to other variances rather than to the candidate's abilities.
- Be aware that tasks should be developed with the target test takers in mind, without favoring or discriminating against test takers who have certain characteristics.
- Be aware that the clarity of writing tasks is essential. Test takers should be able to understand what is required of them quickly and easily.

3.3.4 Writing to CEF Levels

Item writers were required to write items with a difficulty level from B1 to C2 on the CEF scale. Their predictions of item difficulty level would be empirically validated when the items were field-tested. The table below gives an overview of the four main stages in the CEF familiarization trainings for item writers.

The first step provided instruction in some key terms used in the CEF descriptors, and aimed to help item writer trainees understand the CEF. The second step gave trainees a brief idea of the kind of tasks set and how well test candidates were expected to perform at different levels. It also introduced the global descriptors for each level. The third step provided the trainees with more detailed descriptions of 'can do' statements. Table 8 gives an example of CEF overall written production and subscales. Finally, after becoming familiar with the CEF scales, item writers were asked to rate several recordings of speaking performances individually, then to discuss with their colleagues the ratings they had given, and finally to compare their scores and reasons with those given by experts.

Table 7: Item Writer CEF Training Stages

Main Stages	Details
Familiar with the definitions of some basic terms used in CEF	For example: general language competence, communicative language competence, context, conditions and constraints, language activities, language processes, texts, themes, domains, strategies, tasks
Familiar with the common reference level: the global descriptors	**Proficient user (C2 & C1):** precision and ease with the language, naturalness, use of idiomatic expressions and colloquialisms, language used fluently and almost effortlessly, little obvious searching for expression, smoothly flowing, well-structured language **Independent user (B2 & B1):** effective argument, holding one's own, awareness of errors, correcting oneself, maintains interaction and gets across intended meaning, copes flexibly with problems in everyday life **Basic user (A2 & A1):** interacts socially, simple transactions in shops, etc skills uneven, interacts in a simple way
Familiar with the sub-scales for four skills	CEF Overall Written Production and sub-scales CEF Overall Speaking Production and sub-scales CEF Overall Listening Production and sub-scales CEF Overall Reading Production and sub-scales
Rating candidates' performances based on CEF	Rate individually Express reasons and discuss with colleagues Compare rating with experts' marks

As shown in Table 7, there are four steps in the CEF familiarization training. The first step covers the instruction of some key terms that are used in the CEF descriptors, aiming to facilitate all item writer trainees to understand the CEF. By introducing the global descriptors at each level, the second step gave trainees a brief idea of what kind of tasks and how well the test candidates were expected to perform at different levels. It is the third step that provided the trainees with more detailed descriptions of 'can do' statements. Table 8 gives an example of CEF overall written production and subscales. Finally, after becoming familiar with the CEF scales, item writers were asked to rate several recordings of speaking performances individually, discuss with their colleagues what ratings they gave and compare their scores and reasons with those given by experts.

Table 8: An example of CEF Overall Written Production and sub-scales

	CEF Overall Written Production	CEF Writing sub-scales
C2	Can write clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader find significant points.	• Creative writing • Reports and essays • Overall written interaction • Correspondence • Notes, messages and forms • Note taking • Processing text • Orthographical control • Thematic development • Coherence and cohesion
C1	Can write clear, well-structured texts on complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples and rounding off with an appropriate conclusion.	
B2	Can write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesizing and evaluating information and arguments from a number of sources.	
B1	Can write straightforward connected texts on a range of familiar subjects within his/her field of interest, by linking a series of shorter discrete elements in a linear sequence.	
A2	Can write a series of simple phrases and sentences linked with simple connectors like 'and', 'but' and 'because'.	
A1	Can write simple isolated phrases and sentences.	

3.3.5 Monitoring Item Writers' performance

Quality control for item writers continues after the completion of their training. Individual and group performance for both item writing and item reviewing are closely monitored by Test Development staff. The monitoring measures are primarily related to three criteria:

1. Overall, can the item writers develop appropriate items? What percentages of items have been rejected or accepted?
2. What types of item does each writer specialize in or is not good at?
3. What are the reasons for item rejection?

Item writing monitoring procedures include the review of

- Percentages of total number of items submitted with review comments and flagged as discuss
- Percentages of total number of items submitted that are accepted after final Test Development review
- Percentages of items submitted per item type with review comments and flagged as discuss
- Percentages of items submitted per item type that are accepted after final Test Development review
- Quality of items and review comments on a random selection of items.

3.4 Item Development and Item Reviews

3.4.1 Item Development

Item writers were commissioned to write sets of items. Each set contained a maximum of 6 items types. The item writers followed detailed Item Writer Guidelines and the checklists detailed above to ensure that a high proportion of items would be accepted. Item writers used an authoring tool which enabled texts, questions and media to be uploaded to the item bank software automatically. The workflow for PTE Academic items is detailed in Figure 1. This workflow is supported by the item bank software which means that all the stages of item development up to Innovative Item Editor (IIE) migration is controlled and monitored, that all access is tracked, and that all comments on items and decisions are monitored. Information from the item bank software is also fundamental to the feedback given to item writers at a later stage.

The item content review process, immediately following the item writing, is considered another vital part of the validation process. Although the Guidelines highlight the concerns of construct irrelevance and construct under-representation, both may still exist as a result of, for example, the writers' different cultural backgrounds, genders, ages, or religious beliefs. The target test-taking group of PTE Academic is heterogeneous, may speak a wide variety of first languages, will have a wide variety of cultural and social backgrounds, and will come from, or intend to study in a wide variety of academic disciplines. If the content of a task upsets certain groups of test takers, it may affect their test performance. Validity checks by item writers, external content reviewers, and internal Pearson reviewers were therefore conducted to further evaluate the appropriateness of item content and to eliminate potential bias. Figure 1 displays the steps involved in the item development and review process.

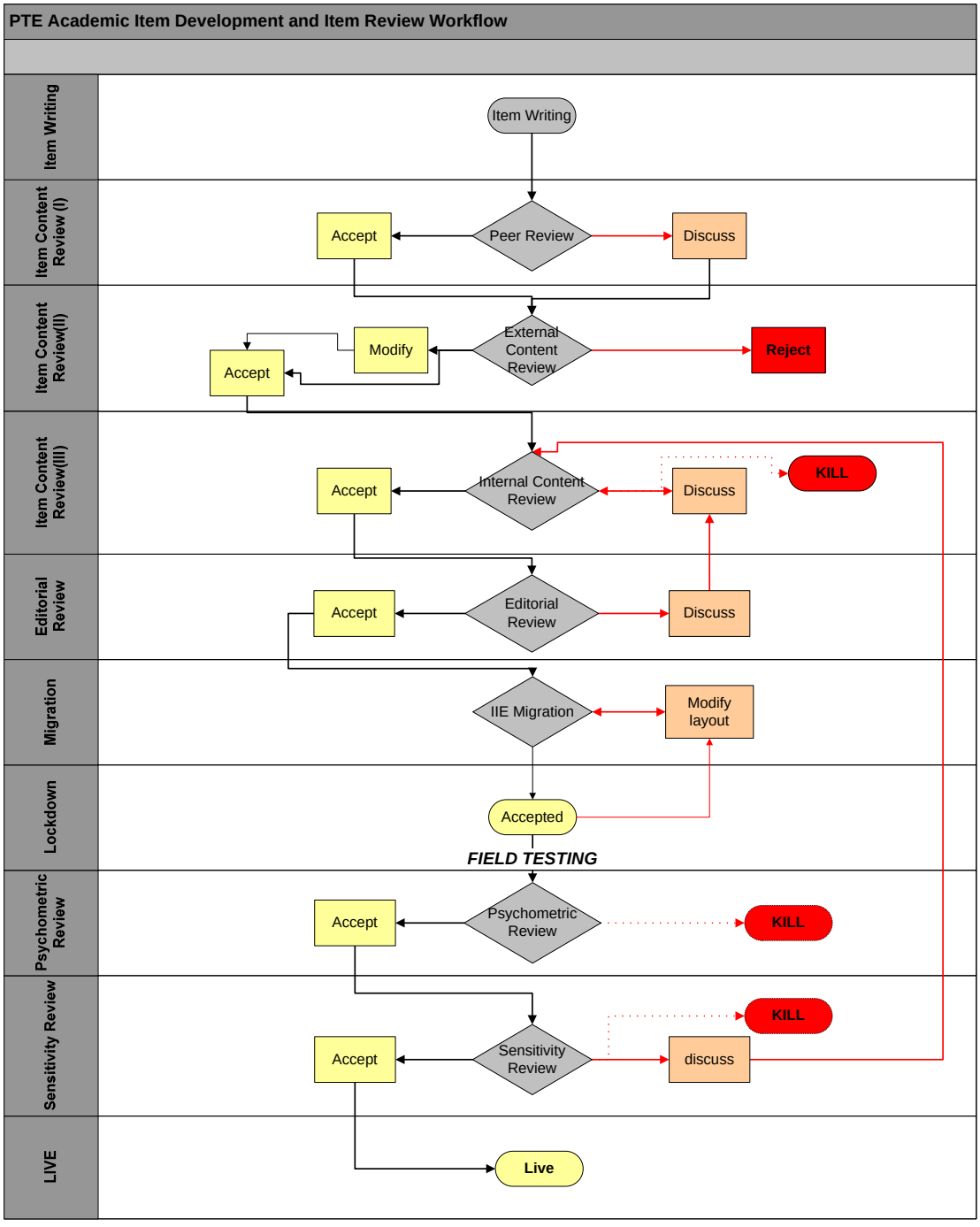

Figure 1: PTE Academic item development and item review workflow[1]

[1] Sensitivity is incorporated into the three-stage item content review process after the field testing.

3.4.2 Item review

In order to mitigate against parochial topics and items, all items written were peer reviewed by item writers from the other two countries. For example, the items written by the UK team were split in half, with one half reviewed by the USA item writers and the other half by the Australia item writers. Peer review processes included checking item dimensions, academic relevance, authenticity, bias, and level of difficulty. Peer reviewers were instructed to focus their review on questions related to authenticity and sensitivity. The relevant parts of their checklists in this context include:

✓ Check for fit to specifications

- Does length of stimulus match specification?
- Does linguistic complexity of stimulus match specification?
- Does genre of stimulus match specification?
- Does rhetorical structure of stimulus match specification?
- Is stimulus appropriate in terms of academic topic?
- Is stimulus example of materials that students are likely to read or listen to in university academic settings?

✓ Check for construct-irrelevant variance

- Could test taker get an item correct from world or background knowledge alone?
- Does item introduce any requirement with respect to subject matter content knowledge?

✓ Check for item difficulty

- Is the CEF level appropriate for the content of each item?
- Is the level of knowledge and skills called for appropriate to the difficulty level?

✓ Check for language, culture, and gender bias

Three types of status were given to the peer reviewed items: 'accept', 'reject', and 'discuss/pending' (see Table 9). If no issue was identified by the peer reviewer, the item was accorded the status of 'accept'. In Field Test One in 2007, the USA item writing team achieved a rejection rate in excess of 32%. Accordingly, extra item writer training was carried out in the USA to ensure that item writers produced higher quality items. In Field Test Two in 2008, all three item writing teams obtained an acceptance rate of over 80%.

Table 9: Peer review results for Field Test One and Two

	Accepted (%)			Rejected (%)			Discuss/Pending (%)		
	UK	USA	AUS	UK	USA	AUS	UK	USA	AUS
2007	87.40	64.50	80.50	11.60	32.00	17.60	1.00	3.60	1.90
2008	83.90	82.60	84.00	10.10	12.20	11.10	6.10	5.10	4.90

Five major reasons triggered item rejection, although not all of them represent threats to construct validity:

- Content: when items demonstrated bias and sensitivity issues or when they failed to meet test specifications; when answers could be obtained based on general knowledge without context, required special knowledge, or were of an inappropriate genre
- *Audio*: when the audio quality was poor, or when further editing was needed
- *Graphic*: when the graphic quality was poor, or when further editing was needed
- *CEF:* when stimulus or questions were out of test range or unsuitable for that level
- *Copyright*: when copyright permission was rejected

After the item writer peer review cycle was completed, external content review and internal content reviews were conducted to further evaluate whether the items were written to measure the intended constructs and nothing else. Items were then set to accept, reject, or modify depending on the results of the review.

3.5 Field Testing

Field test items were operationalized in field tests to gain their psychometrical properties and carry out further concurrent validity studies. In this section, field test demographic information is presented first, followed by information on how results from the two field tests and the follow-up survey and interviews helped further refine the construct validity of PTE Academic.

In total, 10,402 test takers participated in two rounds of field tests. Test takers were provided with incentives to ensure they were motivated enough to put in reasonable efforts in taking the field tests. Field Test One in 2007 had 6,227 test takers, and Field Test Two in 2008 had 4,175 test takers. These test takers were from 158 different countries and spoke 126 different languages. Among the total population of field test takers, 12% were native speakers of English. There were slightly more female test takers than male test takers (54% vs. 46%), even though Field Test One demographics showed slight differences from that of Field Test Two. Overall, the test takers demographics from the two field tests were representative of the target test taker population of PTE Academic.

3.5.1 Feedback from Field Test One

The aim of Field Test One was to understand test takers' feelings about their experiences of taking PTE Academic and to gain feedback . To inform further test item development, semi-structured interviews and a survey were conducted after Field Test One. This section includes an introduction to how the two research instruments were designed, selected findings, and a brief description of the revisions made to achieve better construct validity for PTE Academic.

Design of research instruments

6,221 candidates completed a short survey after Field Test One. The design of the survey primarily took the format of five Likert-scales (strongly agree, agree, no opinion, disagree, and strongly disagree), other than question 1. The instrument was intended to measure a variety of issues regarding test takers' perceptions: overall experience, issues related to computer delivery system, clarity of instructions, and difficulty level of tasks. Details are presented in Table 10 below.

Table 10: Survey questions from Field Test One

	The questions	What the questions measure
1	Please select answer that best describes your opinion	Overall experience
2	The test was easy to navigate	Overall experience
3	It was difficult to know when to start/stop speaking	Device/ instruction
4	I would sit this test again	Overall experience
5	The speaking section was easy	Speaking tasks
6	I would recommend this test to friends	Overall experience
7	The audio was easy to understand	Device
8	I was distracted by others speaking	Test administration
9	The reading section was difficult	Reading tasks
10	The writing section was unfair: I cannot type fast	Writing tasks
11	The listening section was difficult	Listening tasks

Follow-up semi-structured interviews were conducted with 30 test takers covering their concerns about Field Test One in terms of layout, structure, difficulty level, item type and content. These interviews provided supplementary data to the test development team on test takers' perceptions and helped highlight areas for further improvement. The main interview prompts and the number of comments are presented in the table below.

Table 11: Interview prompts and responses after Field Test One

Interview prompts	No. of comments
Overall, how did you find the experience? (structure, layout, difficulty, timings, etc)	15
Did you encounter any significant problems?	13
Do you think the test gives a good measure of your English?	10
Which item type did you find most challenging?	12
Please provide general comments on Reading Question	6
Please provide general comments on Writing Question	8
Please provide general comments on Speaking Question	11
Please provide general comments on Listening Question	11

Findings and recommended changes

Overall, both the interviews and survey data suggested that an overwhelming majority of respondents enjoyed the test experience. The main problems or challenges the test takers encountered which could possibly undermine the argument of establishing construct validity included:

1. Some test takers felt that because it included many different item types and was constructed in a new format (especially for the speaking, listening and integrated tasks), the overall difficulty level of the field test was higher than that of other English tests of a similar nature, such as TOEFL and IETLS. As a result, some items took them much longer to understand and get familiar with.
2. Regarding the recording and speaking device provided by the test centers, 79% of interviewees mentioned distractions caused by other candidates typing on keyboards or performing speaking tasks in the same room.
3. Among those test takers who had problems completing the essay writing, the majority asked for more time because of their weak typing skills.
4. Even though they appreciated the overall clarity of the test instructions and test structure, candidates felt that the standard directive for certain item types could be made more user-friendly.

Revisions were made based on test takers' feedback and on the above findings.

1. A beep was added to the start of speaking types which contain preparation time to signal the opening of the microphone. This aimed to reduce test takers' uncertainty of when to start speaking.
2. New headsets with better sound deadening capabilities, boom quiet headset and Plantronics headsets were piloted in Field Test Two.
3. The timing of the essay was extended from 15 to 20 minutes to write 200-300 words in consideration of the test takers' feedback regarding time pressure.
4. Regarding the clearness of instruction before each type of task, the standard directive for the item type 'select missing word' was made more transparent as the numbers of correct scores for this item was significantly lower. The time for reading the directive was also extended from 7 seconds to 10 seconds.

3.5.2 Feedback from Field Test Two

This section describes how PTE Academic items from Field Test Two were analyzed to ensure satisfactory item statistics, and, consequently, to maintain validity. First, native speakers' responses are reported to gauge the extent to which their inclusion helps support construct validity. Secondly, statistical procedures adopted in the item level analysis, especially criteria developed for item exclusion, are briefly summarized. In total, 6,207 test takers and 1,323 items were analyzed.

Native vs. non-native

Native speaker performance plays an important role in informing item quality (Clark, 1977). In that the intended use of PTE Academic is to determine whether foreign language students have sufficient command of English to participate in tertiary level education where English is the language of instruction and communication, an important step in validation is to ascertain which test items could be answered appropriately by those students who have English as their native language. During field testing, therefore, 10% to 15% of native speakers of comparable age and educational background to the targeted test taker population were included, and the performance of those native speakers constituted one of the item selection criteria.

Native speaker data were analyzed with regard to their response times and their response correctness in order to provide better evidence of validity. In terms of response time, as shown in Table 12, native speakers of English had a median total test time of 113 minutes, 12 minutes faster than that of non-native speakers of English.

Table 12: Test time for native and non-native English speakers

	N	Mean	Median	Min	Max	SD
Native	73	112.66	113	74	161	19.271
Non-Native	843	124.33	125	59	176	19.484

Note: Test time was calculated in minutes.

The timing analysis of all item types was conducted separately. For the three writing item types, the results indicated that for item type 8 and item type 15, where test takers are asked either to summarize a written text or to summarize a spoken text, they took a varying amount of time and only a few test takers used the maximum time. This suggests that the time allowed was sufficient. In contrast, the analysis for item type 17 shows that a large proportion of students took the maximum time allowed on this item type, suggesting that a significant number of test takers would have needed more time to complete these items to their own satisfaction. Consequently, it was decided to lengthen the response time from 15 minutes to 20 minutes for this item type. Similar analyses were carried out for all speaking items. Table 13 shows the descriptive statistics of standardized score (z-score) from both native speakers and non-native speakers in Field Test Two. Native speakers displayed higher scores in total as well as in the four communicative skill scores, but they had a slightly higher standard deviation in speaking.

Table 13: Descriptive statistics of total score and trait scores

	z-score	N	Min	Max	Mean	SD
Native Speakers	Total	198	-1.898	2.792	1.014	0.955
	Read	198	-2.016	2.024	0.786	0.848
	Listen	198	-1.470	2.980	1.066	0.971
	Write	198	-2.290	2.379	1.126	0.706
	Speak	198	-2.039	2.626	0.812	1.259
Non-native Speakers	Total	5507	-3.135	2.945	-0.036	0.979
	Read	5507	-3.683	2.267	-0.028	0.990
	Listen	5507	-2.757	3.139	-0.038	0.976
	Write	5507	-3.359	2.471	-0.040	0.982
	Speak	5507	-2.578	2.808	-0.029	0.974

Classical item statistics were used to support an initial round of item inspection by examining item difficulty, correct keying, and skill specific point-biserials. P-values of all items were calculated. Items were removed when: 1) items had a lower proportion of correct answers from native speakers than from non-native speakers; 2) the non-native p-values were either greater than .90 or lower than .10; 3) item-total correlations were less than or equal to .05; 4) item-total correlation for one or more of the distracters was greater than that of the item-total correlation of the keyed option (rdt>rkt).

Table 14: Removed items by criteria of removal and item type

Item Type	p<.10 or p>.90		rit< 0.05		rdt>rkt		Total Removed	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
01-RR-SAMC	6	11%	0	0%	1	2%	7	13%
03-RR-HOTS	1	2%	0	0%	0	0%	1	2%
04-RR-DRDR	0	0%	1	2%	0	0%	1	2%
08-RW-CONC	0	0%	1	3%	0	0%	1	3%
09-LL-SAMC	5	10%	1	2%	3	6%	9	18%
10-LL-MAMC	1	2%	1	2%	0	0%	2	4%
11-LL-GAPS	1	2%	2	4%	1	2%	4	7%
16-LS-REPT	2	1%	0	0%	0	0%	2	1%
21-LS-SAQS	21	10%	11	5%	0	0%	32	16%
Total items removed	37		17		5		59	
Average removal percentage		4%		2%		1%		7%

The above selection criteria resulted in items being removed from nine item types (see Appendix A for a complete description of item types). Table 14 shows the number of items per item type that were removed according to each criterion. No items had a lower proportion of correct answers from native speakers than from non-native speakers. In total, 59 items were removed for further analysis, with 37 items, 17 items, and 5 items meeting criteria two, three, and four respectively.

Item level analysis

This section presents the psychometric validation procedures endorsed in the field testing periods using the software program ConQuest (Wu, Adams, & Wilson, 1998). The focus is on addressing how one item type was excluded from the test, and how items that have below quality standards were excluded from the item pool to support validity.

The pilot data consisted of large numbers of linked item sets, each administered to a minimum of 200 subjects. Linking of each set with the total set was ensured through a hundred percent overlap, 50% with each of two other item sets. Item sets were carefully balanced to be representative of the total set.

Because of the size of the collected data, no IRT program was available to analyze it in a single run. Therefore, the complete item response dataset was split into two equally sized datasets based on an odd/even item split. A common-examinee linking design was adopted in which an entire collection of 1,318 items was divided into 659-item sets, and a separate calibration was performed for each item set based on the same 5,705 examinees.

A Partial Credit/Rasch model analysis was applied to all odd-numbered items simultaneously. Fit statistics were evaluated according to infit/outfit criteria with misfitting items subsequently deleted. A second analysis was applied using only the even-numbered item dataset, resulting in misfitting items being identified and deleted following the analysis. Common-examinee linking was used to place the even-item parameter estimates on the metric of the odd-item calibration. The even-item calibration was then linked to the odd-item calibration by assuming the mean and variance of the latent trait to be the same across calibrations, that is, the item threshold parameters for the even-item calibration were linearly transformed by applying the same slope (.996) and intercept (+.003) needed to equate the latent trait mean and variance estimates from the even-item calibration to the odd-item calibration. The odd-item calibration was therefore arbitrarily treated as the base metric. The approach necessitated by the size of the dataset in fact had the advantage of allowing a true split-half estimate of reliability.

Table 15 provides a summary of the number of items misfitting according to several criteria. Two ranges of mean square outfit (.8 to 1.2 and .7 to 1.3) and two levels of significance (T>2 and T>3) were investigated. Additionally, due to the large number of misfitting items in item type 03-RR-HOTS, fit statistics for analyses including and excluding that item type were conducted.

Table 15: Misfitting items by item type

Item Type	Fit Criteria including 03-RR-HOTS				Fit Criteria Excluding 03-RR-HOTS			
	.8 to 1.2, T>2	.8 to 1.2, T>3	.7 to 1.3, T>2	.7 to 1.3, T>3	.8 to 1.2, T>2	.8 to 1.2, T>3	.7 to 1.3, T>2	.7 to 1.3, T>3
01-RR-SAMC	1	1			2	2		
02-RR-MAMC	17	11	2	2	25	17	8	8
03-RR-HOTS	49	43	38	38				
04-RR-DRDR	2	1			4	1	1	1
05-RR-GAPS	6	3	3	3	11	6	3	3
06-RL-HILI	1	1	1	1	1	1	1	1
07-RS-READ	1				3			
08-RW-SUMM	13	11	8	8	16	11	9	9
09-LL-SAMC								
10-LL-MAMC	21	14	10	9	28	20	14	13
11-LL-GAPS					1	1		
12-LR-HOTS	23	11	10	10	27	16	17	13
13-LW-GAPS	9	7			6	4		
14-LW-GAPS	15	9			17	7		
15-LW-SUMM	5	3	2	2	5	3	2	2
16-LS-REPT	8	8			8	8		
17-WW-ESSA								
18-RW-GAPS	3	1	1	1	4	3	1	1
19-SS-DESC								
20-LS-PRES								
21-LS-SAQS					1	1		
22-RL-DIAL								

Infit and outfit statistics at the item as well as item type level were inspected in order to identify both individual items and item types that might be considered for deletion. Initial IRT analysis identified a large number of items from item type 03-RR-HOTS with particularly large infit and outfit statistics. To further evaluate the cause of misfit, the empirical versus model-based expected score curves for these items were investigated using Conquest.

The misfit seen for these items is representative of the form of misfit seen for virtually all items of this item type (see Figure 2 for an example). The observed item scores tend to fall below what is expected at the high end of the ability scale, and are also higher than expected at the low end of the ability scale. Such a pattern of residuals is consistent with an item displaying less discrimination than the model implies. This item type was therefore excluded from the test construction pool.

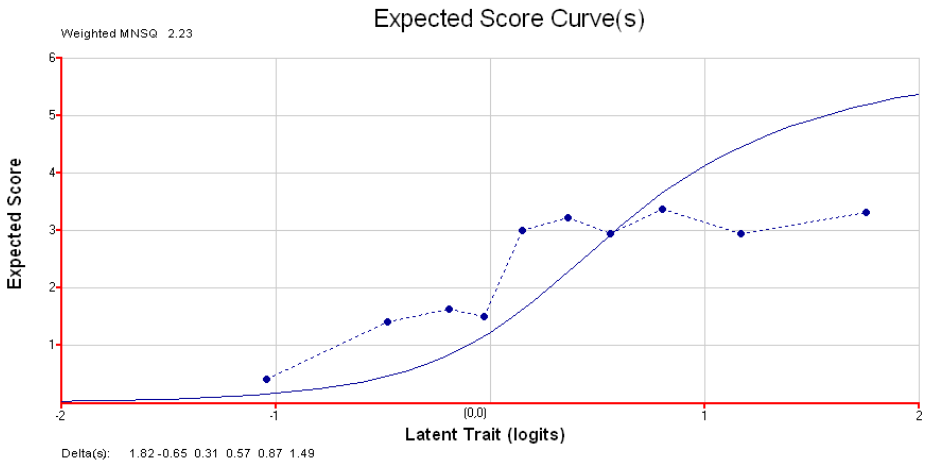

Figure 2: Expected Score Curves for one misfitting Item Type 3 item

Item sensitivity review

In addition to the validity checks by peer reviewers, external content reviewers and Pearson internal reviewers, a three-phase mixed approach item sensitivity review process was conducted to review the contents of the item bank of PTE Academic in order to detect and remedy any instances of bias against, or in favor of, particular groups of test takers. Figure 3 demonstrates the end to end review process.

Phase one was an expert judgment review carried out by a panel of 15 people representing 14 distinct nations and regions. The aim of this phase was to make recommendations on sensitivity to different cultures, religions, ethnic and socio-economic groups, disabilities, gender roles, use of positive language, symbols, words, phrases and content, and on whether an item requires field-specific knowledge. The reviewers were all highly proficient in English and had extensive experience in teaching English as a second/foreign language.

The items were sent to the panelists who were asked to rate them according to a three point scale. An item would be rated as 0 when no sensitivity issues were found, and the item was to be kept in its present form; an item was to be rated 1 where the source of sensitivity was localized within the editable text of the item, such that it could be removed by deleting or substituting a few words. An item would be rated 2 if the source of sensitivity was distributed through the item, or if it was to be found in an audio or video recording, as it was expected to be difficult to edit audio or video material while still maintaining the integrity of the item.

Each item was first reviewed by two panelists, one from the eastern and the other from the western hemisphere. Overall, 83.5% of the items reviewed were rated as 0 by both panelists, were thus deemed to be unproblematic with regard to sensitivity, and were accordingly not subject to further scrutiny. The remaining cases covered items where at least one reviewer had rated 1 or 2. In these cases the Chair examined the item and adjudicated on the reviewers' decisions. For each of

these items, the Chair made recommendations to "keep", where the item was not deemed sensitive and hence could be kept as was; to "edit", where the item was deemed sensitive and hence should be edited so as to remove the sensitivity; or to "kill" where the item was deemed sensitive, could not be edited, and hence should be removed from the item bank.

Table 16 shows the results both of the panelists' ratings and the Chair's adjudications. In this table, the summary of recommendations is a composite of the sum of the panelists' ratings and the Chair's recommendations. For example, "1edit" represents those which one panelist rated as 1 and which the Chair recommended to be edited; "2keep" represents those items which were given a total rating of 2 (either each panelist rated the item as 1 or one rated it as 2 and the other as 0) and the Chair's recommendation was to keep. A list of examples of each category is presented in Appendix B.

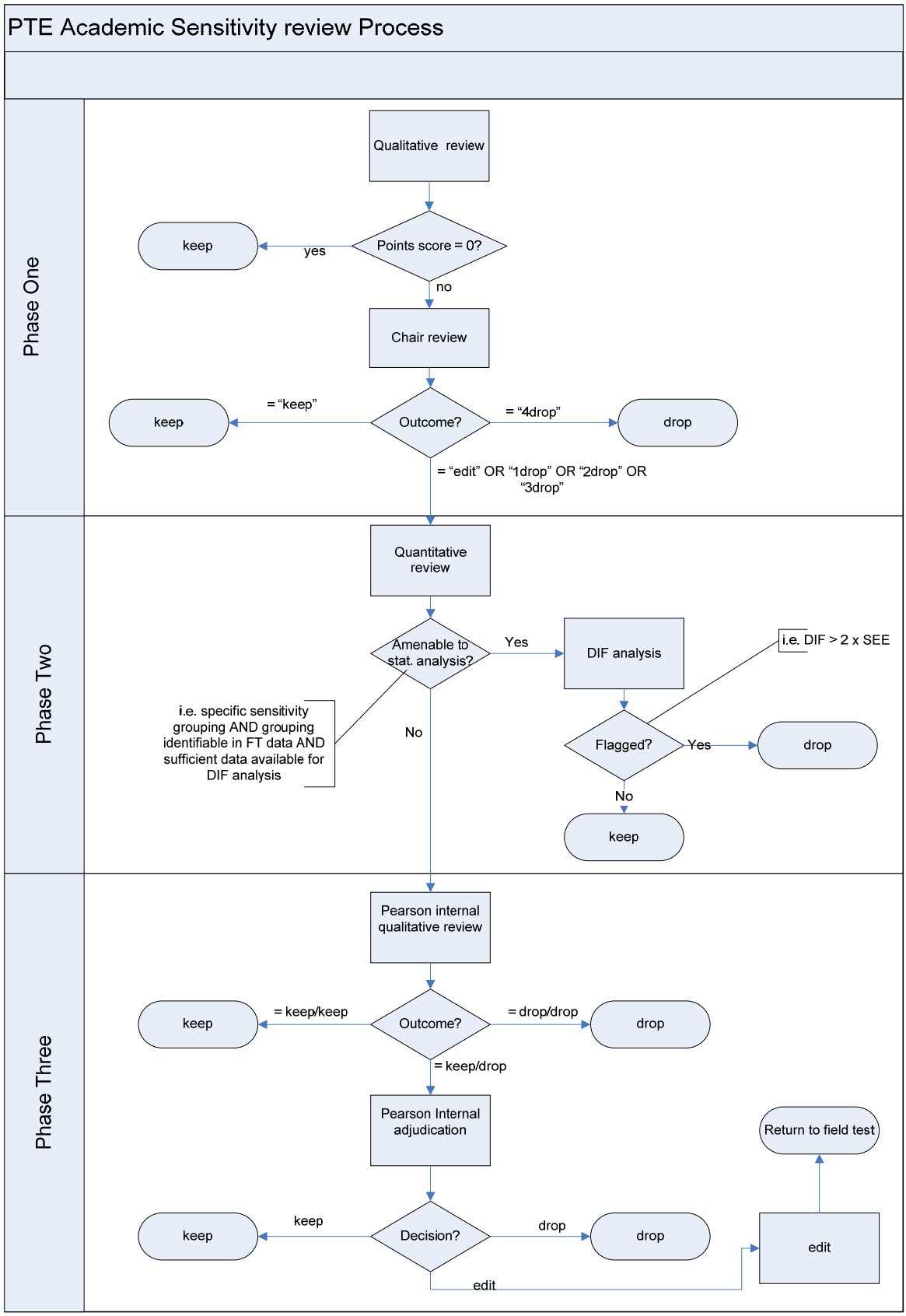

Figure 3 : PTE Academic sensitivity review end to end process

Table 16: Results of panelists' ratings and Chair's adjudications

Summary of recommendations	Percent	Cumulative percent
1edit	15.49	15.49
1keep	37.09	52.58
1kill	8.45	61.03
2edit	3.05	64.08
2keep	13.38	77.46
2kill	15.26	92.72
3edit	0.47	93.19
3keep	1.64	94.84
3kill	3.52	98.36
4keep	0.23	98.59
4kill	1.41	100.00

It can be seen that the most common result is "1keep", meaning that one of the two panelists found some editable sensitivity in the item, but the Chair judged that the item could be kept without editing. In only a very small number of cases (1.41% of adjudicated items) was there a unanimous recommendation that the item should be removed.

Phase Two was a Differential Item Functioning (DIF) analysis with 426 items that were flagged by one or more raters. The 223 items that were deemed not sensitive in the adjudication process were not subjected to further investigation. The remaining 203 items were then examined for suitability for statistical review. A total of 167 of these items were removed from the analysis for various reasons, either because of how they were characterized in the review or because of constraints on the available test data from Field Tests One and Two. The first removal criterion (the manner in which the items were characterized) included items that were identified as general sensitivities, with no groups identified for comparison; items identified as focused sensitivities, with no specific groups specified by the rater/adjudicator; or items with unidentified groups for which there was no useable operational definition of group membership or any suitable proxy variable available for analysis (i.e. developing countries, religion, etc.). Items were also removed due to data constraints. The largest number of items falling into this category was those for which no response data existed in the cleaned combined Field Test One and Field Test two data sets. These items were either removed in previous phases of analysis or may not have been administered in either field test. Additionally, some of groups to which items were flagged as potentially sensitive were present in such low numbers that an analysis would not have been feasible.

A total of 40 items remained for statistical review after the removal process and a limited number of potentially sensitive groups remained. For each of the categories, a dichotomous variable was created for each test-taker that indicated whether they belonged or did not belong to the focal group. The variables used as grouping variables are listed in Table 17, with the total number of test-takers that were classified as belonging to these groups noted in parenthesis.

Table 17: Identified Sensitive Groups that had an Adequate Number of Response for Review

Female (N=3,744)
Region of Birth: Asia (N=6,654)
Region of Birth: Europe (N=1,232)
Region of Birth: Latin America (N=362)
Region of Birth: Asia or Africa (N=6,940)
Region of Birth: Asia, Africa, Latin America (N=7302)
Region of Birth: China (N=1,931)
Field of Study2: Science (N=443)
Field of Study: Humanities (N=314)
Field of Study: Social Sciences (N=323)
Field of Study: Business or Finance (N=1,607)
Field of Study: Education (N=159)

Even after the formation of this shortened list of items for review, substantial limitations were encountered. Typically, a differential item functioning (DIF) approach could be used to assess, after accounting for individual ability, whether people in the specified groups performed differently on a suspect item. There are many methods of assessing DIF, and all require sizable numbers of people in the two groups across which performance is to be compared. While, overall, the size of the focal groups listed above well exceeds the minimum number required to perform an analysis capable of detecting DIF, the actual number of examinees who responded to each item is substantially lower with 50% of the items having fewer than 200 total responses and 80% fewer than 400. When this was coupled with the focal and reference group counts, a large majority of the studied items had sample sizes of fewer than 50 in either the reference or focal group. Additionally, many of the suspect items were polytomously scored, thus requiring even larger samples than would be recommended for dichotomously scored items. While some of the dichotomously scored items form the largest focal groups (i.e. women) had acceptable number of both groups present to employ preferred DIF methods, the great majority of items did not have sufficient sample sizes. In order to keep the methods of review consistent across items, an alternative method, a less preferred method of DIF detection, the delta-plot method, was used to assess sensitivity across the predefined groups.

Although there is some precedent for using delta-plot to assess DIF, this method is most often used in equating designs where each form contains a set of common items. Whether applied to DIF analyses or equating practices, the method is used to identify items that are systematically more difficult for one set of examinees than for another. In general, p-values are computed for each item for the two groups for the comparison of interest. Each set of conditional p-values is then transformed to the delta scale, which has a mean of 13 and a standard deviation of 4, with a linear transformation of the inverse normal equivalent of p-values for each item for each of the two groups. The bivariate delta-plot can then be created with a point for each item, where the delta values for one group on the x-axis and on the y-axis.

If the two comparison groups are of the same ability the points will form a tight ellipse from lower left to upper right. Differences in overall ability (i.e. impact in the context of DIF), will cause this ellipse to shift horizontally or vertically. When other factors impact performance on some items, the points corresponding to the items will fall some distance away on the off diagonals from the ellipse containing most

[2] In each of the Field of Study variables, not all test-takers supplied this information on the demographic survey or supplied information that was not categorized into one of these fields of study. For the purposes of this study, these candidates were not placed into either group, and hence were not used in the review of items that specified any of these groups as potentially sensitive group.

items. The determination of outliers is accomplished by drawing the best linear function to the points, and then calculating the perpendicular distances of each point to the line. The fitted line is chosen to minimize the sum of squared perpendicular distances. This method is unlike ordinary least squares regression, which uses the sum of squared vertical distances to define the line of best-fit i.e. it is a symmetric linear function. The slope of this line is defined by the ratio of standard deviation of all delta values for all items across the two groups, and the intercept defined as the mean of the one group's delta values minus the mean of the other groups delta values that has been weighted by the slope (see equation below).

$$\hat{Y}_{\delta j} = \left(\frac{\sigma_{\delta_2}}{\sigma_{\delta_1}}\right)\delta_{1j} + \left[\bar{\delta}_2 - \left(\frac{\sigma_{\delta_2}}{\sigma_{\delta_1}}\right)\bar{\delta}_1\right]$$

The distance of each point to the best-fit line (i.e. the residual) is then calculated by subtracting the observed δyj from the predicted value of δyj using δxj. The set of residuals can then be used to set some guidelines about what constitutes an item than is unusually difficult for one group relative to all other items behavior across the two groups by defining the standard error of estimation as the standard deviation of all residuals across all items. Outliers are the points with residuals that fall far outside of the typical range (e.g. 2 or 3 SEEs away from the line of prediction), and are items that warrant removal or further investigation.

In this investigation the following steps were taken for each of the 13 comparison groups listed in Table 17.

1. Delta values were computed for all items for each of the two groups on the comparison
2. The line of best fit was estimated as described above, but excluding the items flagged as potentially sensitive to one of the two groups in the comparison.
3. Standard error of estimation was computed using the residuals of items, again excluding the studied items
4. Using the slope and intercepts estimated in step 2, predicted values and residual were computed for the suspect items indentified for the groups in the comparison.
5. Suspect items where the absolute values of the standardized residuals were greater than 2 were flagged.

Table 18 showed the results from the DIF analysis, including grouping information, the delta values for the two groups, and the perpendicular distance from the line of best fit. Four items were identified as statistically sensitive as two items had standardized residuals greater than 2 and two other items had standardized residuals greater than 3.

Table 18: Results from the DIF analysis

Grouping	Delta (Group1)	Delta (Group2)	SE	Residual (Standardized)
Asia/Africa	11.86	13.00	1.07	2.17
Latin America	6.63	13.86	1.25	5.96
Area of study (education)	6.42	11.06	1.58	4.08
Area of study (education)	6.42	9.56	1.58	2.44

An item that was flagged as biased in both the bias-sensitivity review and the DIF analysis used 'America' to refer only to the United States. One of the reviewers from Latin America commented that:

> Although most people in Latin America understand that 'America' and 'American' usually refers to just the United States, the use of these terms remains a serious sensitive cultural issue throughout our countries since, technically, we are all Americans because we all live in the American continent with three regions (North, Central and South), and not as two or three different continents.

Phase Three was another round of item scrutinizing conducted by two Pearson content reviewers to further review the items that were identified in Phase One, but not in Phase Two. Two content reviewers examined the items from the perspective of whether the items posed any bias or sensitivity issues. When the reviewers agreed, the item was kept in the item bank; when the reviewers disagreed, an adjudicator was invited to make recommendations. Overall, 37% of the internally reviewed items were dropped from the item bank. A further 13% were judged to be editable, and were therefore removed from the item bank, edited, and then re-introduced as new items to go through field testing.

This three-phase item sensitivity review greatly helped improve item quality, thereby supporting the construct validity of PTE Academic. One outcome of the sensitivity review was the decision that the sensitivity guidelines be incorporated into all stages of the item development process. It should also be noted that items remaining in the item bank satisfied both statistical as well as content criteria.

Sensitivity review is incorporated into the content review process after the field testing is completed.

4. Supporting evidence for PTE Academic concurrent validity

Concurrent validity refers to the degree to which two different measuring systems produce correlating results. An element of the validity argument for a new test therefore, is to provide evidence of such a relationship with existing measuring instrument that have established and recognized usage for measuring the same or a similar construct.

Concurrent validity evidence was collected during the development of PTE Academic. This section firstly reports the statistical validation procedures used to establish the extent to which PTE Academic scores can be linked to the Common European Framework of Reference for Languages (CEF). The CEF describes what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop to be able to act effectively. Language ability is described with CEF as a number of scales, which include a global scale, skill specific scales, and linguistic competency scales. Statistical procedures for relating PTE Academic scores to the levels of the CEF scales involved both a test taker-centered approach and an item-centered approach. Secondly, this section reports the results from a study of the relationship of PTE Academic with TOEFL and IELTS.

4.1 Linking to CEF: A test taker-centered approach

For the test taker-centered approach, test taker responses on five items from three item types were used: Writing essay (one item); Oral description of an image (two items); and Oral summary of a lecture (two items). Writing essay has 11 scores categories (0-10 points); Oral description of an image has 8 score categories (0-7 points), and Oral summary of a lecture has 5 score categories (0-4 points). These responses were rated on the relevant CEF scales for writing and speaking by two human raters, independently of the ratings produced to score the test. Given the probabilistic and continuous nature of the CEF scale, adjacent scores were expected in the model.

The relation between ability estimates based on scored responses on the above PTE Academic test items and the CEF is displayed in Figure 4, with one for the written responses, and the other for the oral responses. The horizontal axis ranges from CEF levels A2 to C2. The vertical axis shows the truncated PTE Academic theta scale varying from -2 to +2. The box plots show substantial overlap across adjacent CEF categories, as well as an apparent ceiling effect at C2 for writing. CEF levels, however, are not to be interpreted as mutually exclusive categories. Language development is continuous, and does not take place in stages. Therefore, the CEF scale and its levels should be interpreted as probabilistic: learners of a language are estimated most likely to be at a particular level, but this does not reduce to zero their probability to be at an adjacent level.

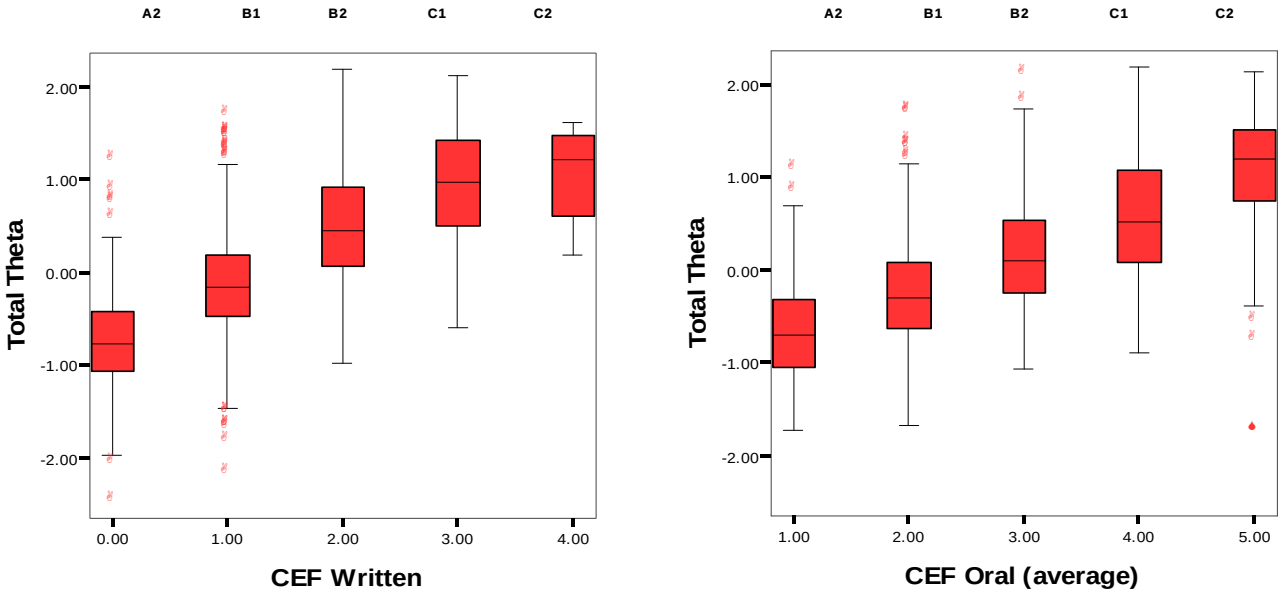

Figure 4: CEF level distribution Box Plots

Although the official CEF literature does not provide information on the minimum probability required to be at a CEF level, the original scaling of the levels (North, 2000) is based on the Rasch model where cut-offs are defined at 0.5 probability. The distance of approximate 1 logit between levels implies that anyone typically reaching a probability of around 0.8 at level X, has 0.5 probability of being at level X+1 and is therefore exiting level X and entering level X+1. Having a probability of 0.5 of being at level X implies a probability of 0.15 to be at level X+1 and as little as 0.05 at level X+2. Based on the monotone increase of the PTE Academic theta from

A2 to C2 as shown in Figure 6, a positive relation between the CEF scale and the PTE Academic scale is established. To find the exact cut-offs on the PTE theta scale corresponding to the CEF levels, the first step is to establish the lower bounds of the CEF categories based on the independent CEF ratings. For this purpose, the CEF ratings were scaled using FACETS (Linacre, 1988; 2005).The estimates of category boundaries on the CEF theta scale are shown in Table 19.

Table 19: Category lower bounds on CEF theta

Category	CEF level	CEF theta (Lower bounds)
0	BELOW a2	N/A
1	A2	-4.24
2	B1	-1.53
3	B2	0.63
4	C1	2.07
5	C2	3.07

The relationship between the scale underlying the CEF levels and the PTE Academic theta for those test takers about whom we had information on both scales (n=3,318) is shown in Figure 5. The horizontal axis is the CEF theta, and the vertical axis is the PTE Academic theta estimate. The correlation between the two measures is 0.69. A better fitting regression is obtained with a first order polynomial (curved red line), yielding an r2 of slightly over 0.5. This regression function was used to project the CEF cut-offs from the CEF scaled ratings onto the PTE Academic theta scale.

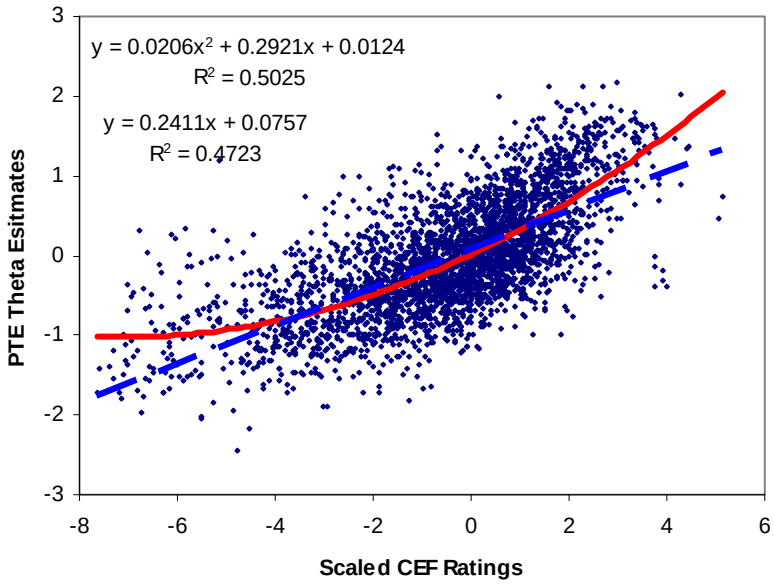

Figure 5: Relation between CEF theta and PTE theta

Because of noisy data at the bottom end of the scales, the lowest performing 50 candidates were removed. Further analyses were conducted with the remaining 3,268 subjects. Figure 6 shows the cumulative frequencies for these 3,268 candidates for whom theta estimates are available on both scales (CEF scale and PTE Academic scale). The cumulative frequencies are closely aligned though the PTE scale clearly shows smaller variance.

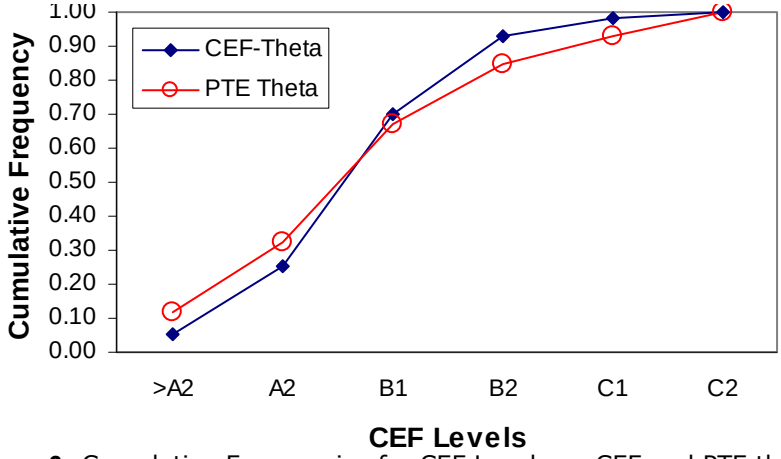

Figure 6: Cumulative Frequencies for CEF Levels on CEF and PTE theta scales

In the next stage, an equipercentile equating was chosen to express the CEF lower bounds on the PTE theta scale. Equipercentile equating determines the equating relationship as one where a score has an equivalent percentile on either form. The cumulative frequencies are shown in Figure 7 and the projection of the CEF lower bounds on the PTE theta scale together with the observed distribution of field test candidates over the CEF levels is shown in Table 20.

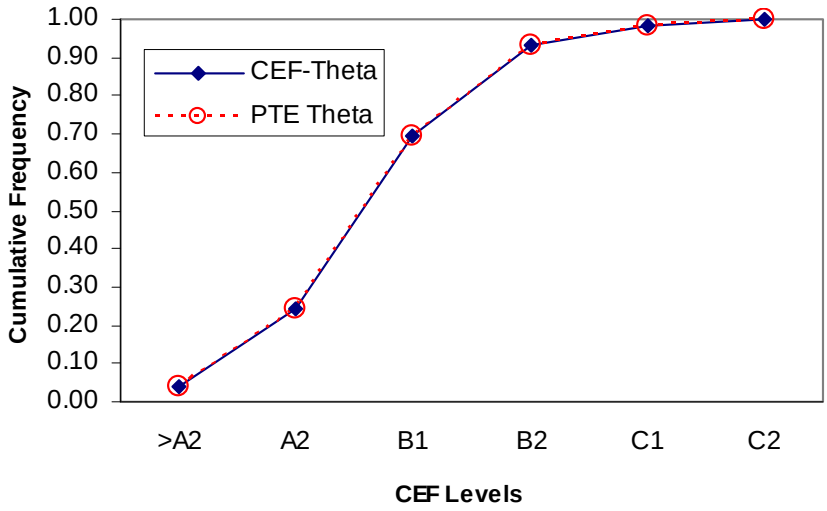

Figure 7: Cumulative frequencies on CEF and PTE theta scales after equipercentile equating

Table 20: Final estimates for CEF lower bounds on PTE theta scale

CEF Levels	Theta PTE	Frequency	Percentage	Cumulative Frequency
<A2	-1.366	126	4%	0.04
A2	-1.155	677	21%	0.25
B1	-0.496	1471	45%	0.70
B2	0.274	769	24%	0.93
C1	1.105	170	5%	0.98
C2	>1.554	55	2%	1.00
	TOTALS	3268	100%	

4.2 Linking with CEF: An item-centered Approach

At the item development stage, item writers were required to indicate for each item which level of ability expressed in terms of the CEF levels they intended to measure, i.e., did they think test takers would need to be able to correctly solve the items. In the item review process, these initial estimates from item writers were evaluated, and if needed, corrected by the item reviewers. Based on observations from field tests, the average item difficulty was calculated for items to fall into a particular category according to item writers. Table 21 provides the mean observed difficulty for each of the CEF levels targeted by the item writers.

Table 21: Intended and observed item difficulty

Intended CEF Level	Mean observed difficulty
A2	0.172
B1	0.368
B2	0.823
C1	1.039
C2	1.323

However, the cut-offs on the PTE Academic theta scale need to be established based on item writer estimates. To this effect, from the data, given item difficulty, the likelihood of any item to have been assigned to any of the CEF levels was estimated. The cut-offs between the two consecutive levels is the location on the scale where the likelihood of belonging to the first category becomes less than the likelihood of belonging to the next category. In this way, the PTE theta cut-offs based on the items were found. The estimated lower bounds of the difficulty of items targeted at each of the CEF levels were plotted against the lower bounds of these levels as estimated from the independent CEF ratings of test takers' responses by human raters. In Figure 8, the horizontal axis represents the CEF cut-offs from the test taker-centered analysis, while the vertical axis represents the CEF cut-offs from the item-centered analysis. Both estimates, derived independently, agree to a high degree (r=0.99).

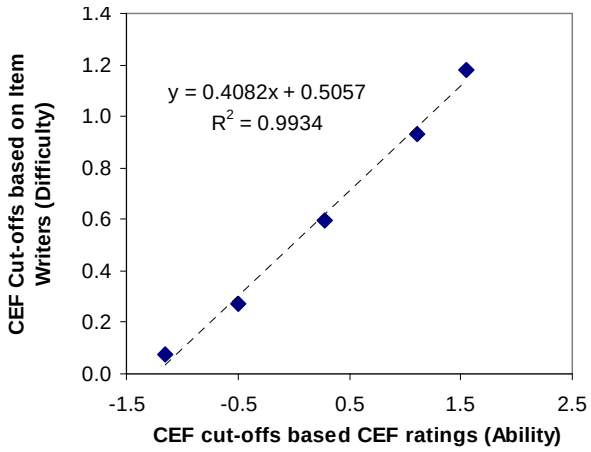

Figure 8: Lower bounds of CEF levels based on targeted item difficulty versus lower bounds based on Equated CEF ratings of candidates' responses

4.3 Concordance with other measures of English language competencies

A concordance study between PTE Academic and other measures of English language competencies was conducted during the field testing stage. Test-takers self-reported scores on other tests of English, including TOEIC, TOEFL PBT, TOEFL CBT, TOEFL iBT and IELTS. In addition, test takers were asked to send in a copy of their score reports from these tests. About one in four of all test takers that provided self-reported scores also sent in their official report. Table 17 indicates that the correlation between the self-reported results and the official score reports was .82 for TOEFL iBT and .89 for IELTS. This finding is in agreement with earlier research on self-reported data. For example, Cassady (2001) found students' self-reported GPA scores to be 'remarkably similar' to official records. The data are also consistent. According to ETS (2005, p. 7) the score range 75-95 on TOEFL iBT is comparable to the score range 213-240 on TOEFL CBT and to the score range 550-587 on TOEFL PBT. Table 22 shows the mean of the self-reported scores in those tests and their corresponding correlation with PTE Academic.

Table 22: Means and correlations of PTE Academic field test takers on other tests

Test	Self-reported data			Official score report		
	n (valid)	Mean	Correlation	n	Mean	Correlation
TOEIC	327	831.55	.76	na		
TOEFL PBT	92	572.3	.64	na		
TOEFL CBT	107	240.5	.46	na		
TOEFL iBT	140	92.9	.75	19	92.1	.95
IELTS	2432	6.49	.76	169	6.61	.73

In addition, according to ETS (2001, p.3) a score range of 800-850 on TOEIC corresponds to a score range of 569-588 on TOEFL PBT, which also makes the self-reported TOEIC mean scores of the test takers on the PTE Academic field test fall in line with data published by ETS.

Table 23: Correlation and prediction of PTE Academic BETA test takers

Test	Self-reported data				Official score report			
	n	Mean	Predicted	Correlation	n	Mean	Predicted	Correlation
TOEFL iBT	42	98.9	97.3	.75	13	92.2	98.2	.77
IELTS	57	6.80	6.75	.73	15	6.60	6.51	.83

Based on the data presented in Table 23, concordance coefficients were generated between PTE Academic and other tests of English using linear regression. The regression coefficients were then used to predict the scores of PTE Academic BETA test takers' scores on TOEFL iBT and IELTS. Table 23 shows the self-reported mean scores and those from the official reports, the mean scores from the same test takers as predicted from their PTE Academic score, and the correlations between the reported scores and predictions from PTE Academic.

Combining the results from concordance with CEF as well as concordance with other English tests, two complete concordance tables have been generated based on the established conversion coefficients, one among PTE Academic, TOEFL iBT scores, and CEF, the other among PTE Academic, IELTS, and CEF.

5. Conclusions

This paper has presented only part of the growing body of work which supports the validity claims of PTE Academic. Examples of other validation work, outside the scope of this paper, include the validation of machine scoring and lexical validation. This paper represents an attempt to collect in one document the wide range of work, both qualitative and analytical, which has contributed and continues to contribute to the development of high quality test items. This included a carefully planned test specification, scheduled item writer training and reviews, and a comprehensive, phased field test program involving over 10,000 participants.

This paper has reported on the measures taken to support the construct validity of PTE Academic. The results demonstrate that the varying of item types, presentation modes, content of stimuli, task types, required competences, and response format has enabled the construction of a nomothetic network, resulting in a coherent descriptive model of test takers' English language ability by mode of language use. Furthermore, a global perspective and standard has been developed from the organized processes involved in item writing and item peer review and a mixed-method approach for item sensitivity review has provided us with a systematic method of scrutinizing items for potential bias and/or sensitivity. Data from two field tests both underpins and permits the comprehensive analysis presented in this paper. Pearson Language Tests is committed to collecting more evidence. Concurrent validity has also been established from the beginning of the test development process by mapping onto the CEF and by comparing results from other tests of a similar nature. This work will continue alongside other standard setting requirements.

Given that PTE Academic is a relatively new test, studies of many other aspects of validity need to be carried out in further research projects in order to fulfill the requirements set by Standards for Educational and Psychological Testing (AERA, APA & NCME, 1999). The standards propose five validity resources: (a) evidence based on test content, (b) evidence based on response processes, (c) evidence based on internal structure, (d) evidence based on relationships to other variables, and (e) evidence based on consequences of testing. There is a funded program in place to support research into these issues and to sponsor other critical validity studies, such as predictive and consequential validity. The work done to date on validating PTE Academic continues to be scrutinized by an external Technical Advisory Group which adds further weight to the validity proposition for PTE Academic.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly*, *25*(4), 671-704

Cassady, Jerrell C. (2001) Self-Reported GPA and SAT Scores. ERIC Digest. ERIC Identifier: ED458216.

Cizek, G. J. Rosenberg, S. L. & Koons, H. H. (2008). Sources of Validity Evidence for Educational and Psychological Tests. *Educational and Psychological Measurement*, *68* (3), 397-412

Clark, J. (1977). *TOEFL research reports: The performance of native speakers of English on the test of English as a foreign language*. Princeton: ETS

Council of Europe (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge: Cambridge University Press.

ETS (2001) TOEFL Institutional Testing Program (ITP) and TOEIC Institutional Program (IP): Two On-Site Testing Tools from ETS at a Glance. Handout Berlin Conference 2001. Princeton: Educational Testing Service.

ETS (2005) TOEFL ® Internet-based test: Score comparison tables. Princeton: Educational Testing Service.

Linacre, J.M (1988; 2005) A Computer Program for the Analysis of Multi-Faceted Data. Chicago, IL: Mesa Press.

Messick, S. (1989). Validity. In R. Linn (ed.), Educational Measurement (pp.13-103). New York: Macmillan.

Messick, S. (1992). Validity of test interpretation and use. In M.C. Alkin (ed.), *Encyclopedia of Educational Research* (6th ed.). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741-9

Moller, A.D. (1982).A Study in the Validation of Proficiency Tests of English as a Foreign Language.Unpublished PhD thesis.University of Edinburgh.

Schilling, S. G. (2004). Conceptualizing the validity argument: An alternative approach. *Measurement*, *2*, 178-182.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. New York: MacMillan.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). Modelling polytomously scored items with the rating scale and partial credit models. *In ACER ConQuest: Generalized item response modelling software* (pp. 25-37). Melbourne, Australia: Australian Council for Educational Research.

Appendix A: PTE Academic item type description

Item type	Brief Description	Task Description
01 RR SAMC	Multiple choice, choose single answer	Test takers need to read the text and select a single answer.
02 RR MAMC	Multiple choice, choose multiple answers	Test takers need to read the text, and select all response options that apply.
04 RR DRDR	Re-order paragraphs	The stimulus presents 4 or 5 randomly ordered sentences. Test takers need to reconstruct the text by moving the sentences and placing them in a comprehensible and coherent order.
05 RR GAPS	Reading: Fill in the blanks	The stimulus presents a gapped real-life reading text. After understanding the meaning of the text and reading the alternatives carefully, test takers select from the alternatives the word or phrase that best completes each gap in the text.
06 LR HILI	Highlight correct summary	The stimulus presents a real-life audio/video of an academic lecture or speech. The stimulus also presents 3-5 paragraphs. After listening to the audio/video, test takers select the correct paragraph.
07 SR READ	Read aloud	The stimulus presents a short real-life reading text of 40-60 words. Test takers need to read the text aloud once.
08 RW SUMM	Summarize written text	The stimulus presents a reading text. Test takers need to read the text and summarize it using one sentence of up to 30 words.
09 LL SAMC	Multiple choice, choose single answer	The stimulus presents an audio or video recording about an academic subject. Test takers need to listen to the recording and select a single answer.
10 LL MAMC	Multiple choice, choose multiple answers	The stimulus presents an audio or video recording about an academic subject. Test takers need to listen to the recording and select all response options that apply.

11 LL GAPS	Listening: Fill in the blanks	The stimulus presents an audio or video recording about an academic subject. The last word or group of words in the passage is replaced by a short electronic beep. Test takers need to listen to the audio recording and choose the option that best completes the audio text.
12 LR HOTS	Highlight incorrect words	The stimulus presents a real-life, authentic recording. The stimulus also presents a reading text, which is a transcription of the audio recording containing 5-6 deliberate "errors". While listening to the audio, test takers click on all the words that differ from what they have heard. Selected words are highlighted after the test taker has clicked on them.
13 LW GAPS	Select missing word	The stimulus presents a real-life, authentic audio excerpted from an academic lecture/speech, or a conversation typical of those that occur on a university campus. The stimulus also presents a reading text which is a transcription of the audio recording with 4-7 words missing from the text. Test-takers need to listen to the audio and complete the gapped written text by typing the missing word in each gap.
14 LW DICT	Write from dictation	The stimulus presents a short sentence of 8-11 words. Whilst listening to the audio, test takers transcribe what is spoken and type the exact sentence in the space provided.
15 LW SUMM	Summarize spoken text	The stimulus presents a real-life, authentic audio/video excerpt from an academic lecture. Test takers need to listen to the audio recording, and write a summary of what the speaker has said.
16 LS REPT	Repeat sentence	The stimulus presents a short scripted recording. After hearing the sentence, test takers repeat the sentence exactly as they hear it.
17 WW ESSA	Write essay	The written prompt consists of 1-2 sentences (30 - 50 words) that instruct test takers to express their views on a general academic topic. Test takers write a persuasive essay and support their position or opinions with details and examples.
18 RW GAPS	Reading & writing: Fill in the blanks	The stimulus presents a real-life gapped reading text from an academic source. There are 4-5 gaps in the text. Each gap has a drop-down list with 4 possible choices to complete the gap. Test takers need to complete the gapped text by selecting a word from the drop-down lists for each gap.

19 SS DESC	Describe image	The stimulus presents one or more images (e.g. graph, picture, map, chart, and table) from an academic source. After looking at the image(s) on full screen, test takers describe in detail the development or sequence of events presented graphically.
20 LS PRES	Re-tell lecture	The stimulus presents a real-life, authentic audio/video excerpt from an academic lecture together with a visual such as a PowerPoint presentation or similar media to enhance understanding. Test-takers hear an audio recording/watch a video, and are to retell what they have just heard/watched in their own words.
21 LS SAQS	Answer short question	The stimulus presents a short spoken question, which asks for basic information, or requires simple inferences. Test takers answer the question with a single word or a short phrase.

Note: item type 03-RR-HOTS is dropped during the field test

Appendix B: Examples of comments from panelists and chair in sensitivity review

Results	Item type and controversial points	Panelists' comments	Chair's comments
1 (0+1) edit	Reading passage Offensive to certain group of people	Journalists might be offended by this sentence: Journalists do not need to present the same rigorous referencing and support for their claims as social scientists	Delete "We need to remember, though, that journalists do not need to present the same rigorous referencing and support for their claims as social scientist are required to do."
1 (0+1) keep	Reading passage Content of the reading passage may favor those test takers who have priori knowledge about computer technology	This may be biased for the students who have learnt these steps already and it is also culture specific.	Computer dialogue menus are pervasive, world wide, and particularly so in academic work (the target test-taking population). As to whether or not the students have already learned the steps, the item is detailed enough that even if a student happens to have been in an academic course with exactly this sequence of events, the nature of the key and distracters force him/her to read closely. I see no advantage to prior experience.
1 (0+1) kill	Listening passage Content of the listening text is likely to trigger offensive attitude from test takers with certain background	This passage is about an acerbic remark on education. It is better tone down the more outspoken passages such as they are way underpaid, undereducated….etc.	This passage runs the risk of causing a negative emotional reaction among test-takers who are teachers or who have worked in education.
2 (1+1) edit	Reading passage May be offensive to conservative readers	Panelist 1: Change the clause …" Much like an ordinary woman …. candlelight" and also change the second option Panelist 2: The reference to women looking especially beautiful by night might be a little bit too 'forward' for a more conservative reader,	I agree with both reviewers. Edit both the passage and (probably change) the option to remove discussion of the physical beauty of women.

		as it indirectly talks about sexuality and discusses it openly. Note the same reference in the answer as well. Instead, it might be more appropriate to say something like 'individuals have a better appearance in candlelight".	
2 (0+2) keep	Listening passage The tone or attitudes of the speaker may or may not emotionally affect the test takers	A very complicated text and with negative perspective towards people. I have a feeling that the speaker finds all the people dangerous and distrustful. (one of the panelists rated 2)	I disagree with panelist 1. I think the speaker is contextualizing his remarks (yet to come). The entire segment reads like a kind of normal getting-started hedge.
2 (1+1) kill	Reading passage The topic of 'maternal death' and the attitudes towards it from various religions and culture may have culture bias	Panelist 1: It is better not to mention the two countries, which reported material deaths but did not report deaths due to abortion. What if we change the phrase like "Tow countries, located in Northern part of Asia and East Asia. Panelist 2: Wording of option 4 could change from 'religious' to 'social or customary'	Quite apart from mentioning specific countries (which, clearly, could be rectified by editing), the overall topic of the passage is sensitive: maternal death. Test-takers with experiences related to maternal death (e.g. their own mother died in childbirth) may have a very negative reaction to this task.
3 (1+2) edit	Reading passage The content may be sensitive to some professions and values behind the content may have bias to people from some	Panelist 1: According to the passage, among children farmed out to wet-nurses, the odds were much worse - up to two thirds died. It might give a negative impression of wet-nurses	Change "among children farmed out to wet-nurse" to "among children under care by wet-nurses". Try also to remove the mention of Angola -- perhaps remove the entire sentence. Also change 'replicate western values' in option 02 (the first option listed) to 'replicate historical values' -- that may make it both a better distracter and also remove its potential sensitivity.

	countries	on society even though it is statistically verified. Panelist 2: wording of the first option "western values" need to change	
3 (1+2) keep	Reading passage Content of the reading passage may favor those test takers who have priori knowledge about science	Panelist 1: This may be a difficult item for non-science test takers. (reasons to delete) Panelist 2: As above, this question is based on prior knowledge about a light year rather than on English.	The particular scientific knowledge here is sufficiently common that I doubt it will raise a sensitive reaction.
3 (1+2) kill	Reading passage Content of the reading materials may insult test candidates from certain country	Panelist 1: This item presents distributed sensitivity, since throughout the text it operates by stereotyping about Brazilian women's asserted body image and its recent changes. In a way it can be insulting for candidates with cultural or social links with the country. Panelist 2: The first sentence is a little bit stereotyping. The removal of the first sentence may probably reduce the sensitivity. In the first option 02, there is one sentence: Brazilian women tended to be fatter than the American or European ideal. In my opinion, Brazilian women tend to	clearly disturbing to test-takers who have grappled with the issue of body image

		think that the above sentence is a little bit offensive.	
4 (2+2) keep	Reading passage Some words may or may not favor a group of test takers with background knowledge of geography	Panelist 1: the same as above, here testing geographical knowledge so biased for the students of geography and similar disciplines. Panelist 2: Completely technical question which relates to general knowledge and not English	The particular scientific terminology (really only the word 'hemisphere') here is sufficiently common that I doubt it will raise a sensitive reaction.
4 (2+2) kill	Reading passage The topic of 'violence' may be offensive to some test takers with similar experiences	Panelist 1: The topic is very sensitive; candidates who have experienced violence might get upset or distressed. Please, check the recording, the end of the last sentence is missing. Panelist 2: This is a distributed sensitivity. Suppose the test takers have experienced this violence before, and then the process of taking this test will be a bad experience for them. It could only bring back many bad memories. The item, in my point of view, should be completely removed.	I agree with both reviewers