Research Note: Applying EALTA Guidelines – A Practical case study on Pearson Test of English Academic

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

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

March 2011

1. Introduction

Similar to the fields of educational testing and psychological testing, standards, guidelines, and codes of practices in the field of language testing are prolific and they serve the purpose of guiding different language testing industries to have baseline values in the tests they produce. Internationally, there is ILTA guidelines for practice by International Language Testing Association (2007). Regionally, to name a few, there are ALTE code of practice by Association of Language Testers in Europe (1994), ALTE principles of good practice for ALTE by Association of Language Testers in Europe (2001), and JLTA code of good testing practices by Japanese Language Testing Association (2002). There are also standards produced by individual testing organizations, for example, ETS standards for quality and fairness Educational Testing Service (2002).

Despite the abundance of these standards and guidelines, how they are observed in practical testing developments is rarely documented; furthermore, there have observed an even sparse application of these standards or guidelines in practical test development practices. This article focused on providing a review of the application of the European Association for Language Testing and Assessment (EALTA) Guidelines for Good Practice in Language Testing and Assessment (EALTA, 2006) to the development of a new international language test, Pearson Test of English Academic (PTE Academic).

According to its mission statement, the purpose of the EALTA is to promote the understanding of the theoretical principles of language testing and assessment and the improvement and sharing of testing and assessment practices throughout Europe. One of the instruments by which EALTA pursues its goals is through the publication of the Guidelines for Good Practice in Language Testing and Assessment (EALTA, 2006). The EALTA guidelines are available in more than thirty languages and were developed in order to provide general principles guiding good practice in language testing and assessment. In the course of developing a new language test, it is, therefore, appropriate and useful to verify whether and wherever relevant the EALTA guidelines are observed. At the same time, to examine the advantages and disadvantages of applying guidelines like the EALTA guidelines in real-life tests, the results of which may have high-stakes consequences.

The EALTA guidelines for good practice in testing and assessment are targeted at three different types of audiences: 1) those engaged in the training of teachers in testing and assessment; 2) classroom testing and assessment; and 3) the development of tests in national or institutional testing units or centers. Focusing on the guidelines targeted at the third type of audience, the test development process of PTE Academic was checked against the seven critical aspects as defined by the

EALTA Guidelines: 1) Test Purpose and Specification; 2) Test Design and Item Writing; 3) Quality Control and Test Analyses; 4) Test Administration; 5) Review; 6) Washback; and 7) Linkage to the Common European Framework (CEFR). The purpose of this article is to show how Pearson strives to adhere to the principles of transparency, accountability and quality appropriate to the development of PTE Academic, and to enhance the quality of the assessment system and practice. Empirical research on the EALTA guidelines mainly includes Alderson (2010) and Alderson & Banerjee (2008). They devised their survey questionnaire to the aviation English tests providers on the above seven aspects. Relating to the use of codes of practice, ethics, or guidelines for good practices, the authors argued that guidelines, such as the EALTA guidelines could be used to 'frame a validity study' (Alderson, 2010, p. 63).

2. PTE Academic in the Context of the EALTA Guidelines

The following sections are organized in the order of the seven aforementioned aspects. Answers to the questions are listed under the subheadings below. Specific examples, documents and the ways the guidelines have been observed are summarized within each section.

2.1 Test purpose and specification

This section presents the test purpose of PTE Academic and how the test specification was used in the test development process.

How clearly is/are test purpose(s) specified?

The purpose of PTE Academic is to accurately measure the communicative English language skills of international students in an academic environment. The test requires test takers to engage in a wide range of interactive and integrative tasks based on live samples of English language use in academic settings. The primary use of PTE Academic is to make decisions about students' readiness to study at English-medium educational institutions. The test purpose is clearly stated in the test specification document.

How is potential test misuse addressed?

To avoid potential misuse of the test, detailed information on how to appropriately interpret and use PTE Academic test scores is provided in three documents available on the PTE website Interpreting the PTE Academic Score Report, Using PTE Academic Scores, and Skills and Scoring in PTE Academic. Additional materials such as the Standard Setting Kit are also available to aid score users in setting standards for using scores at their institution for admission purposes.

Are all stakeholders specifically identified?

Test stakeholders are identified to be test takers and test score users, the latter group including universities, higher education institutions, teachers, government departments and professional associations requiring academic-level English. The stakeholders are clearly described in the test specification document.

Are there test specifications?

Once decisions had been made about the purpose of the test, the domains and construct that was to be measured and the intended use of the test, the test development team designed the test by creating detailed test specifications. The specifications delineate the test purpose, constructs, framework of the instrument, test length, context in which the instrument is to be used, characteristics of intended participants, psychometric properties, conditions and procedures for administering the instrument, procedures for scoring, and reporting of the test results. The test specifications have gone through multiple revisions in response to feedback from various sources. A Technical Advisory Group comprising experts from both language testing and psychometrics provided feedback, advice and critical assessment on the test specifications.

Are the specifications for the various audiences differentiated?

The test specifications are used to guide the development of PTE Academic test items and their associated scoring rubrics and procedures. The test specifications have been adapted for various audiences including test takers, test score users, and external researchers. For example, an adapted version of the test specifications is used in the Official Guide to the PTE Academic. An adapted version of the specifications is also available in the form of FAQs for test takers and score users.

Is there a description of the test taker?

The population for which PTE Academic is appropriate is specified to be non-native English speakers who need to provide evidence of their academic English language proficiency, because they intend to study in countries where English is the language of instruction. The target test population is clearly described in the test specification document.

Are the constructs intended to underlie the test/subtest(s) specified?

The construct that PTE Academic is intended to assess is communicative language skills for reception, production and interaction in the oral and written modes as these skills are needed to successfully follow courses and actively participate in tertiary level education where English is the language of instruction. The construct is clearly stated in the test specification document.

Are test methods/tasks described and exemplified?

There are a variety of selected-response item types (e.g. multiple-choice, hotspots, highlight, drag & drop, and fill in the blanks) for assessing the oral and written receptive skills, and a variety of open constructed-response items (e.g. short-answer and extended discourse) for the oral and written productive skills. Each item type is described and exemplified in materials such as the Item Writer Guidelines, the Test Tutorial, and *The Official Guide to PTE Academic.*

Is the range of student performances described and exemplified?

To help clarify the scoring criteria, a range of sample student spoken and written performances at different CEFR levels are described and exemplified in documents *The Official Guide to PTE Academic*, *PTE Academic Score Interpretation Guide,* and *Standard Setting Kit*.

Are marking schemes/rating criteria described?

The marking schemes/rating criteria for each item type are described in documents such as *The Official Guide to PTE Academic.* The analytic procedures for scoring extended-responses are also described in the document *PTE Academic Scoring Rubrics*. The process for test scoring is described in the document *PTE Academic Overall Scoring.*

2.2 Test design and item writing

This section describes how the EALTA standards were applied to the test design and item writing processes.

Do test developers and item writers have relevant experience of teaching at the level the assessment is aimed at?

Three groups of item writers based in the UK, Australia and the US were recruited to develop items for PTE Academic. Most of the item writers have varieties of experience in EFL/ESL teaching and assessment. Each group of item writers is guided by managers highly qualified in (English) language testing.

What training do test developers and item writers have?

Training sessions were conducted before item writing session began. Each item writer received extensive training on how to interpret and use the CEFR, the meaning of the CEFR levels, how to choose appropriate test materials, and how to construct test items that can potentially discriminate between test takers with varying English language proficiency. Additional support was provided throughout the item writing process.

Are there guidelines for test design and item writing?

Detailed item writing guidelines were provided to each item writer. General item writing guidelines focused on general item writing principles (e.g. validity and reliability, authenticity, sensitivity and bias check). Specific item writing guidelines provided detailed advice and guidance on how to select materials and construct items for each of the 20 item types in the test. Procedures for scoring, and scoring criteria for each item type are also presented to the item writers to maximize their understanding of the potential impact of items on test scores.

Are there systematic procedures for review, revision and editing of items and tasks to ensure that they match the test specifications and comply with item writer guidelines?

To ascertain the quality of the draft test items, systematic procedures were adopted to review, revise and edit the items. The peer review process, which immediately followed the item writing process, helped ensure that international English was adequately represented without undue idiosyncrasies of any of the varieties of English. International English in the context of PTE Academic is defined as English as it is spoken internationally by users of English who wish to be easily understood by most other users of English. To do so, item writers from Australia checked items submitted by the UK and US writers. Item writers from the UK evaluated items submitted by Australian and the US writers. Item writers from the US reviewed items submitted by Australian and the UK writers. The peer reviewers had a large amount of input into the test development process. If deemed necessary, they edited the item to make it conform better to the test specifications and item writing guidelines. They were also asked to record their revisions to the items. Items which they felt did not fully fulfill the requirements were flagged as “Discuss” status.

After the peer review process, each item went through a content review process, in which one or more expert judges looked at each item to check for content quality and clarity. The expert judges are made up of a panel of 15 people representing 14 distinct national and regions. Revisions and suggestions provided by peer item reviewers were analyzed and evaluated. Items flagged as “Discuss” status in the previous stage received special attention at this stage. The expert judges were also asked to use the PTE Academic Sensitivity & Bias Review Guidelines to identify materials and items that are likely to be inappropriate, offensive or confusing for certain groups in the test taking population.

The content review process was followed by an editorial review process, in which each item was further checked for content quality, accuracy and clarity. Following the editorial review process, each approved test item was authored in the Innovative Item Editor, a program which creates the interactive computer screens used by the test driver to present the test, and was then again reviewed to ensure content quality, and layout and format consistency.

What feedback do item writers receive on their work?

The total item bank was filled by several round of item writing. At the end of each round, the item writing teams received a detailed evaluation report, which provided feedback and statistical data on the quality of test items produced, both at the level of the separate country ‘teams’ and of the individual item writer. This document aims to bring to light general problems and areas of improvement to be addressed before commencing the next round of item writing.

2.3 Quality control and test analyses

This section addresses quality control measures undertaken for PTE Academic and describes test analysis procedures.

What quality control procedures are applied?

The quality of the items is ascertained through rigorous item review process and large-scale pilot testing. In addition, a Technical Advisory Group, which comprises experts from both language testing and psychometrics, meets regularly to provide feedback, advice and critical assessment.

Are the tests piloted?

Two large-scale field tests involving a total of 10,400 subjects were carried out, one in 2007 and a second in 2008. A beta test was carried out in 2009. All newly developed items will be used as seeding items in live tests and go through an evaluation stage to ensure their quality in content as well as their psychometric properties.

What is the normal size of the pilot sample, and how does it compare with the test population?

10,400 test takers from 21 countries worldwide participated in the two field tests and the beta test. The test takers from these pilot tests resembled PTE Academic test population in three aspects. First, the participating countries include 1) China, Japan, India and South Korea which produce the largest population of English as second language learners; 2) USA, UK, and Australia which receive the largest population of English as second language learners. Second, the target age for PTE Academic is 17 to 35 years of age and more than 92% of PTE Academic field tests participants fell into this range. Third, the test takers from field tests were mostly university students who are the majority PTE Academic test population.

What information is collected during piloting?

At the field test stage, test taker responses to the test items, their detailed demographic information, and their English learning history were collected, along with surveys collecting test takers' and teachers' perceptions of the test. In addition several rounds of in-depth focus group sessions were organized with test takers in key countries.

How is pilot data analyzed?

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. Classical item statistics were used to support an initial round of item inspection, basically looking at difficulty, correct keying, and skill specific point-biserials. 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 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

identified and deleted following the analysis. 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 and intercept 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.

Based on the initial analyses some item types were rescored using alternative scoring models to improve item characteristics and obtain better overall model fit. The rescored items were calibrated along with a large segment of the remaining previously calibrated items. The resulting estimates were then linked to the original base calibration using the same approach described above for the odd-/even-item linking. Similar procedures will be applied when adding new items that have gone through the seeding process.

Test takers' demographic information, their English learning history, as well as test takers' and teachers' perceptions of the test were analyzed to help 1) improve the test design and test formats, 2) evaluate test item quality and difficulty, and 3) understand test taker performance.

How are changes to the test agreed upon after the analyses of the evidence collected in the pilot?

Based on classical test theory, items were removed when: 1) items had a lower proportion of correct answers from native speakers than from non-native speakers; 2) item difficulties were out of the desirable range for non-native speakers; 3) item-total correlations were too low (<.20); 4) item-total correlation for one or more of the distracters was greater than that of the item-total correlation of the keyed option. Based on the results obtained from the above analyses, further item scaling and dimensionality analyses were conducted.

As an ongoing quality control measure, native speakers' responses are included for seeded items within live test forms. The patterns and characteristics of native speakers' responses are to be analyzed in depth on their specific psychometric features to ensure test item quality.

Statistical analyses of pilot test data informed changes to the item selection. For example, one experimental item type out of 21 trialled failed to yield reliable results and was subsequently removed. The psychometric analysis of PTE Academic field test data and Beta test data accomplished the following:

- helped determine scoring models;
- provided data to be used for training and validating intelligent automated scoring systems;
- identified items of substandard quality, which were then eliminated;
- established how item scores can be combined to generate reporting scores;
- established the minimum number of items for each item type in the test structure;
- defined difficulty parameters for all items.

Possible future modification to the test will undergo the same procedures and controls implemented for developing the initial version. Proposals for modification will be submitted to the Technical Advisory group. Data will be collected and submitted to psychometric analyses. Depending on the intention of the modification – whether it affects the test construct - procedures for guaranteeing the equivalence of the scores will be implemented and new score interpretation guidance will be developed.

If there are different versions of the test (e.g., year by year) how is the equivalence verified?

Test forms are continuously constructed through stratified random sampling from the item bank. Draws are stratified to ensure comparable composition with respect to the item types included and thereby the representation of skills in the test. To ensure equivalence of test forms test forms an sampling algorithm is applied that produces maximally equivalent test information functions across all forms for each of the four communicative skills. The target information function ensures a maximum standard error of measurement of 2.7 score points on the scoring scale in the most relevant area for admission decisions. Towards the extremes of the scoring scale the measurement error gradually increases to a maximum of 4 score points.

Are markers trained for each test administration?

In principle, human rating is only used on new items in order to train automated scoring machines. During live testing human rating is used 1) when the error component of the automatic score exceeds a predetermined value of the size that it could jeopardize the precision of measurement of the reported scores; 2) for occasional probing to check the automated scoring; 3) for validation purposes; 4) on seeded items. Human rater standardization was carried out for each round of rating during field testing. Raters had face-to-face rating training and standardization at the beginning of the rating sessions. Before they started the actual rating, they were required to pass a paper-based or computer-based standardization exam. Standardization sessions will again be conducted when new ratings sessions are needed.

Are benchmarked performances used in the training?

Detailed scoring rubrics and examples of test takers' responses that illustrate each score in the scoring rubrics are used in the training session.

Is there routine double marking for subjectively marked tests? Is inter and intrarater reliability calculated?

Written and spoken responses are machine scored using intelligent scoring engines, so effectively there is no subjective marking and double marking would make no sense. The scoring engines were trained using 2.6 million ratings provided by trained human raters of candidate responses that had been scored. In this process double marking was conducted throughout.

Items were rated on several traits independently. Each item trait was scored at least twice and adjudicated when disagreement between raters occurred. The standardization guide that was used for rater training contained sample responses for each rubric score point. These samples had been independently marked and benchmarked by language testing professionals from Pearson and Second Language Testing Inc. High levels of interrater reliability (+/-.90) were achieved and reported to the Technical Advisory Group and at conferences.

Is the marking routinely monitored?

The aim was for the raters to reach 80% exact agreement. Raters with the lowest 10% rater agreement were not commissioned to do any live marking. In addition, the raters were constantly monitored by their supervisors using backreading (a procedure by which supervisors are randomly assigned responses to double check ratings. In addition PKT, the Pearson unit responsible for the automated scoring, conducted background checking. The proportion of background checking increased when rater variation was observed.

What statistical analyses are used?

Both Classical Test Theory (CTT) and Item Response Theory (IRT) are employed to analyze test item data. CTT analyses provide p-values, item-total correlation, maximum scores, mean scores, point-biserial statistics, multiple-choice option statistics. IRT analyses provide item parameter estimates, fit statistics, ability and item difficulty estimates. In addition, a variety of other statistical analyses, including cluster analysis, factor analysis, multiple regression, were used to help understand the underlying constructs measured as well as students' performance profiles.

What results are reported? How? To whom?

PTE Academic employs a detailed, profiled score reporting system. Test takers receive scores in a variety of dimensions: an overall score, scores for each of the four communicative skills, (i.e., Listening, Speaking, Reading, and Writing), and scores for each of six enabling skills, i.e. Grammar, Oral Fluency, Pronunciation, Spelling, Vocabulary, and Written Discourse.

Two versions of score reports are provided: the test taker version and the institution version. Test takers are notified by email when their scores are available, typically within five business days from their test date, for minimal delay in admission applications. Test takers are able to access their PTE Academic scores online with a secure login and password. Recognizing institutions and organizations have access to an enhanced reporting and results service which supports the admission process and assists institutions in making more informed admission decisions. In addition test level data are reported to the external Technical Advisory Group.

What processes are in place for test takers to make complaints or seek reassessments?

If a test taker is unhappy with his/her test score, he/she can request a rescore. Full details of how to proceed with the rescore are provided in the Test Taker Handbook. If a test taker believes there was an error in any part of the test that may have affected his/her score, he/she can complete and return the Item Challenge Form, which allows the test taker to describe the item challenge on any item content, including typographical (spelling) error, problem with item layout, sensitivity complaint, bias complaint, audio quality, and graphic/image quality

2.4 Test administration

This section focuses on the procedures adopted to ensure the test security of PTE Academic.

What are the security arrangements?

Test security is essential to maintaining test integrity. PTE Academic is administered using advanced technologies to ensure maximum test security and to enhance the testing experience. Stringent security measures are implemented within the Pearson VUE test center network, to protect test content and verify the identity of the test taker. Test center staff has no access to item content. In the test center the photograph for the Score report is taken at the time of testing. Government-issued photo-bearing ID is required. Biometrics further includes a palm-vein print and a voice print. Test taker autographs are captured and digitally stored. To ensure the security of test content, the item bank is replenished continuously with items of the appropriate level of difficulty. Randomized test forms are provided to minimize item exposure and possibility of cheating and fraud.

Are test administrators trained?

As PTE Academic is a large-scale testing program, strict administrative procedures need to be followed. A self-paced PowerPoint presentation is available on the Pearson VUE Support Services website for training test administrators. Once training is completed, Pearson VUE requires that the administrator take and pass a no-cost, informal, open-book test on a Pearson VUE Test Delivery workstation.

Is the test administration monitored?

Pearson uses state-of-the-art biometrics to ensure the security of the testing process for PTE Academic. These include using digital photographs, palm vein printing, and digitally stored signatures to authenticate test takers. Test center administrators also implement video and audio monitoring in the test center to provide the highest level of test security. Test takers are seated in partitioned testing seats and cannot see other test taker's screens; they are required to leave all materials that might help them solve items in lockers outside of the test room, and any potentially fraudulent incidents will be thoroughly investigated. In addition, test takers are required to provide a Personal Introduction in the Speaking section of PTE Academic. This provides an additional biometric identification (voice print) to allow institutions to compare an applicant's voice with the voice recorded while taking the test. The Personal Introduction is available as a digital sound file and accompanies the electronic PTE Academic score reports available for institutions on a secured web-application.

Is there an examiner's report each year or each administration?

Since PTE Academic is a computer-based automatic scored test, examiner's report is not applicable in this context. However, test statistics are provided on an ongoing basis to the Technical Advisory Group.

2.5 Review

How often are the tests reviewed and revised?

A large item bank has been developed for PTE Academic. Through stratified random item selection, a quasi-infinite number of unique test forms can be drawn from the item bank. The different test forms are equated by using a program that composes the tests to all match the same target test information functions at the level of the whole test as well as for the four language skills thereby ensuring that different versions of PTE Academic yield ability estimates that can be used interchangeably even through they are based on different set of test items. Test takers' performances on different test forms are constantly reviewed by the psychometric team to check whether the testing and scoring standards are met.

A test taker's survey, which aims to collect meaningful information and feedback on PTE Academic, is being carried out. A comprehensive research program is also being carried out to evaluate the test use and consequences. It is intended that the test and its supporting documents (e.g. *The Official Guide*, test taker handbook, teacher handbook etc.) will be reviewed periodically to determine whether amendments, and revisions are necessary.

Are validation studies conducted?

Test validation, the process of gathering evidence on whether the test is fulfilling its purpose, is ongoing. For PTE Academic, the validation process began with the conceptualization and design of the test. As 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 whether students who have English as their native language can easily obtain scores at or above the score that is set as a minimum requirement for university admission. During field testing therefore all samples included 10 -15 percent of native speakers of comparable age and educational background as the target population of foreign students and native speaker performance on the items constituted one of the item selection criteria.

Other validation efforts during test development included externally conducted studies on the academic level of vocabulary both in the item prompts as in the test takers' responses and on potential bias introduced through item content. During live testing validation studies include test item checking, test score validation and automated scoring process validation. The first refers to the checking of proper functioning of all the items in each of the live test forms. Any test form that contains defective or dis-functioning items will be removed from live testing. Score validation refers to validation carried out to examine how the scoring rules are observed in live tests. Automated scoring process validation comprise two types of procedure: 1) validation procedures adopted to validate human raters for the purpose of training automated scoring machines; and 2) validation procedures applying part of the human marks to validate machine scores. Test validation continues with a growing program of validation research as the test is being used to make decisions about test takers' academic English language proficiency.

What procedures are in place to ensure that the test keeps pace with changes in the curriculum?

Since PTE Academic assesses functional English language competence in the context of English medium higher education, it is not based on any curriculum.

2.6 Washback

This section illustrates how PTE Academic aims to promote positive test washback and to maximize the positive consequence of using the test.

Is the test intended to initiate change(s) in the current practice?

Ensuring a positive washback is a major concern in the design and development of PTE Academic. To make sure that item types and test content are likely to be naturalistic examples of academic tasks and materials, the test design process began with the analysis of the academic tasks and identification of important characteristics of the tasks that can be captured in PTE Academic item types. A number of innovative features have been integrated into the design and development of PTE Academic. For example, PTE Academic employs a variety of innovative item formats that require the integration of two or more language skills. It is hoped that the innovative feature will bring positive changes to the English teaching and learning practice that would put more emphasis on the development of communicative English language skills. By requiring item writers to use real life material and to name their source when submitting items, Pearson can guarantee and prove that the language in PTE Academic is the language that students will indeed encounter when they go to university. Also, by setting tasks that require dealing with this kind of language, the test stimulates students to perform on such tasks with texts they can themselves gather from similar sources.

What is the washback effect? What studies have been conducted?

Washback is defined as the influence of testing on teaching and learning (Hughes, 1989; Alderson and Wall, 1993; Bailey, 1996). It is too early to assess the long-term impact of PTE Academic on teaching and learning, however some initial studies of washback have been conducted. Three focus groups were conducted in June and July 2008 with 23 field test takers. The objective of the focus groups was to interact with field test participants to better understand their attitudes, thoughts, and ideas on PTE Academic. In general, the reactions of students and teachers to PTE Academic were very positive. Below are some sample comments from the participants. The comments suggest the potential positive washback of PTE Academic on English language learning and teaching.

- *The most useful preparation method for this new test should be through real-life practice, instead of learning strategies from language test training schools. More authentic materials would be required to prepare for the test.*
- *PTE Academic gives a better indication of English language abilities. I like the fact that PTE Academic has a lot of new item types and uses real lectures in the test. I felt I was in a classroom when I listened to the audio recording of the lecture retelling task.*
- *The best way to get a good score on PTE Academic will be to improve my overall language abilities. I would put emphasis on the integration of the different language skills.*

Studying washback is a long-term endeavor, and the Test Development Team in will further evaluate the consequences the test might bring about in English teaching and learning through a longitudinal research program to investigate washback of PTE Academic in varieties of aspects.

Are there preparatory materials?

A variety of learning resources, including the tutorial, test taker handbook, sample test, *The Official Guide*, practice tests have been developed and are available for PTE Academic test takers. In addition the website for PTE Academic offers so-called skills Pods for learners. The website links the skills and subskills tested in PTE Academic with learning resources selected from internationally published ELT skills course books (and related online content).

Are teachers trained to prepare their students for the test / exam?

To help teacher prepare their students for the test, a variety of teaching resources have been recommended for teachers. The official test website links the skills and subskills tested in PTE Academic with teaching resources selected from internationally published ELT skills course books and related online content.

2.7 Linkage to the Common European Framework

This section describes how PTE Academic has been linked to the Common European Framework of Reference for Languages (CEFR).

What evidence is there of the quality of the process followed to link tests and examinations to the Common European Framework?

The preliminary relation of the PTE Academic score scale with the descriptive scale of the CEFR is based on both an item-centered and a test taker-centered method. For the item-centered method the CEFR levels of all items was estimated by item writers, reviewed and, if necessary, adapted in the item-reviewing process. For the test taker-centered method, three extended responses (one written and two spoken) per test taker were each rated by two independent, trained raters. On disagreement between the two independent raters, a third rating was gathered and the two closest ratings were retained. A dataset of over 26,000 ratings (by test takers, by items, by raters) on up to 100 different items was analyzed using the computer program FACETS (Linacre, 1988; 2005). Estimates of the lower bounds of the CEFR levels based on the item-centered method correlated at .996 with those based on the test taker-centered method.

Have the procedures recommended in the Manual and the Reference Supplement been applied appropriately?

To ensure that PTE Academic is a valid and accurate measure of English language ability, Pearson has followed the procedures recommended in the Manual throughout the linking process.

As a starting point, the test specifications were developed in agreement with the CEFR framework. Each item writer received specific training in using the CEFR, and they were then asked to consider the CEFR as the construct model for the test design and provide their estimate of difficulty in terms of the CEFR levels of each item they submitted. Through these activities, they gained detailed knowledge of the CEFR through extensive training. These procedures ensured that the definition and production of the test have been undertaken carefully, following good practice outlined in the Manual. The test development team also collected and analyzed empirical test data in order to provide evidence for the linking to the CEFR.

Is there a publicly available report on the linking process?

The report on the preliminary estimates of concordance between PTE Academic and the CEFR is available at the following location to the general public:

http://pearsonpte.com/SiteCollectionDocuments/PremliminaryEstimatesofConcordanceUS.pdf

3. Conclusion

In the field of language test development, as the EALTA guidelines point out, it is important to provide answers to the questions concerned with the seven aspects of the test development. In order to enhance the quality of language assessment systems and practices, it is also important for the test developers to engage in the dialogue with decision makers in the institutions and ministries to ensure that decision makers are aware of both good and bad practice.

Reviewing the application of EALTA Guidelines in the development of PTE Academic has served a dual purpose: on one hand, this review verified that the development of PTE Academic has been undertaken in accordance with internationally recognized standards of good practice in language testing and assessment. On the other hand, the review served as a practical case study in the application of the EALTA Guidelines to the development of a large-scale and high-stakes international English language test.

The result of this review indicated that although very useful in checking against the guidelines in the course of the test development, it probably is not the ultimate tool to decide the quality of these aspects of the test. Alderson & Banerjee (2008) also pointed out that guidelines are good for consciousness-raising and useful in framing validity studies. The guidelines, however, seem to have a lack of certain aspects involved in the current language testing practices and would probably benefit from updating so as to better perform its function in guiding and enhancing the quality of language assessment systems and practices. For example, item banking is receiving great interest in the field, many tests, including PTE Academic, adopt item bank as a repertoire of items, therefore, and an addition of how item banks are developed and maintained would be very relevant in the EALTA guidelines. Similarly, equipped with new technologies, automated scoring, despite of all the criticism it's receiving, is undoubtedly leading the trend of making language testing a 21st business. Guidelines need to cover these new areas to be able to make it relevant and useful.

In addition, language tests take different forms and have different stakes. The EALTA guidelines can be expanded in raising consciousness in better and fairer practices in high-stakes tests. For example, by adding questions regarding addressing needs from different stake-holders. One final reflection is that scientific means of score reporting have been receiving increasing attention in the field (Goodman & Hambleton, 2004; Hambleton, 2010). Making test score reports more understandable and user-friendly is crucial in communicating the test results to stakeholders. Additions on guidelines in providing understandable and user-friendly score reports would assist test providers in producing value-added testing products.

References

Alderson, J. C. (2010). A survey of aviation English tests. Language Testing, 27. 51-72.

Alderson, J. C. & Banerjee, J. (2008). EALTA's Guidelines for Good Practice: A test of implementation. Paper presented at the 5th Annual Conference of the European Association for Language Testing and Assessment, Athens, Greece, May. http://www.ealta.eu.org/conference/2009/program.htm, last accessed 17th August 2010.

Alderson, J. C. & Wall, D. (1993). Does Washback Exist? Applied Linguistics,14(2),115-129.

Association of Language Testers in Europe. (1994). ALTE code of practice.

Association of Language Testers in Europe. (2001). ALTE principles of good practice for ALTE examinations.

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. Language Testing, 13, 257-279.

Educational Testing Service. (2002). ETS standards for quality and fairness. Princeton, NJ: ETS.

European Association for Language Testing and Assessment (2006). Guidelines for Good Practice in Language Testing and Assessment.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. Applied Measurement in Education, 17, 145-220.

Hambleton, R. (2010). A new challenge: Making test score reports more understandable and useful. Paper presented at the 7th Conference of the International Test Commission. Hong Kong,

Hughes, A. (1989). Testing for Language Teachers. Cambridge: Cambridge University Press.

International Language Testing Association. (2007). ILTA guidelines for practice.

Interpreting the PTE Academic Score Report. http://pearsonpte.com/SiteCollectionDocuments/6813_UK_interpreting_PTEASRlr_131009_V4.pdf.

Item Challenge Form http://pearsonpte.com/SiteCollectionDocuments/PTEAcademicItemChallengeForm.doc.

Japanese Language Testing Association. (2002). JLTA code of good testing practices.

Official Guide.http://pearsonpte.com/Testme/Pages/OfficialGuidetoPTEAcademic.aspx

Linacre, J. M. (1998; 2005). A Computer Program for the Analysis of Multi-Faceted Data. Chicago, IL: Mesa Press.

Preliminary Estimates of Concordance between PTE Academic and other Measures of English Language Competencies. http://pearsonpte.com/PTEAcademic/scores/Documents/PTEA_Preliminary_Estimates_of_Concordance_4Aug10_v2.pdf.

Skills and Scoring in PTE Academic. http://pearsonpte.com/PTEAcademic/scores/Documents/Skills_and_Scoring_in_PTEA_4Aug10_v2.pdf.

Test Taker Handbook. http://pearsonpte.com/SiteCollectionDocuments/PTEAcademicTestTakerHandbook.pdf

Using PTE Academic Scores, http://pearsonpte.com/PTEAcademic/scores/Documents/Using_PTE_Academic_v2_13July.pdf