Diagnostic Testing of Hong Kong Tertiary Students’ English Language Proficiency: The Development and Validation of DELTA

Alan Urmston, Michelle Raquel and Carrie Tsang
English Language Centre
The Hong Kong Polytechnic University

Abstract

The Diagnostic English Language Tracking Assessment (DELTA) is a web-based English diagnostic testing and tracking system in use at three universities in Hong Kong. Using Rasch measurement techniques, test items are calibrated on a measurement scale and thereby determine learners’ English proficiency empirically. Rasch measurement techniques also allow for the generation of itemized diagnostic reports that students receive after taking the DELTA, which enables them to better manage their language learning by setting themselves clear goals and targets. The system also allows students to track their English language development as each time they take the Assessment they receive a combination of texts and items that are targeted to their proficiency level. This paper reports on the development of the DELTA as an instrument that assesses Hong Kong tertiary students’ English proficiency and on the results of the initial administration of the DELTA in 2011–12.

Keywords: assessment; diagnostic; proficiency; tertiary; tracking

The Need for a Diagnostic Assessment for University Undergraduates in Hong Kong

Hong Kong has been described as having an examination-oriented education system (Cheng, 1997; Fullilove, 1992), with the pressure to obtain one of the limited numbers of places available at university leading to a culture in which success is measured by performance in the secondary school matriculation examination, the Hong Kong Diploma of Secondary Education. Once they begin university, students are faced with four years of study, usually through the medium of English, which for most of them is a second language. At university, undergraduates of both degree and sub-degree programmes are provided with English enhancement in the form of taught, credit and non-credit bearing courses, self-access learning facilities, extra-curricular and other activities. The nature of the enhancement that students receive varies from institution to institution as well as from faculty to faculty, department to department and even programme to programme within each institution.

As numerous as the English language enhancement activities are the assessments that students need to take. These assessments include course-embedded tests and assessments; assessments designed for placement into English enhancement streams or courses; assessments designed to measure achievement after a certain period of time; and assessments designed to measure proficiency at graduation. In addition to these self-initiated and individualised assessments is the one common
The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA

The development and validation of DELTA
development experience of the Centre’s staff. In late 2010, the Hong Kong Institute of Education withdrew from full collaboration on the project and was replaced by the English Language Centre of the City University of Hong Kong.

During the development of the TELT, the development team found that a diagnostic language test would benefit students more than a proficiency test because its results could be used for formative purposes. However, diagnostic language testing was an area of language testing that has been under-researched, as summarized by Alderson (2007). Although diagnostic testing has been found to be common in second language learning contexts, it was more prevalent in the primary and secondary sectors than the tertiary and was in the main confined to the classroom. There were few, if any, examples of systematic testing of Hong Kong tertiary-level students for diagnostic purposes, and none, as far as the researchers could determine, which made use of Item Response Theory (IRT) or Rasch measurement to calibrate test items and thereby determine learners’ English language proficiency empirically.³

In response to this need, pilot tests were conducted in September 2009 on the existing versions of the TELT on volunteer students studying in Year 1 at the respective universities. This time, results were disseminated to the test-takers in the form of a diagnostic score report. Students were advised to make use of the report to help them with their English enhancement and teachers were also briefed on how to help students understand the report and advise them on suitable actions to take. As a follow-up to the pilot testing, the test-takers’ feedback on the test and on the pilot diagnostic report was sought; the findings indicate that the students found the report quite helpful (Tsang, forthcoming). The report was thus revised accordingly with further enhancement.

Another round of pilot tests was carried out in September 2010, on a new cohort of Year 1 students and on a number of those who had taken part in 2009. Results were disseminated with an enhanced diagnostic report and those who had retaken the test were also given a track graph showing the progress that they had made from their first attempt. Again, feedback from teachers, students, and management on the revised test and report was sought. The results clearly indicate that TELT as a diagnostic test was more useful for teaching and learning. To take into account its diagnostic potential and its ability to track students’ proficiency development if taken on a regular basis, the test was renamed the Diagnostic English Language Tracking Assessment or DELTA.

The Diagnostic English Language Tracking Assessment (DELTA)

The DELTA is based on the Bachman and Palmer (1996) framework of language competence and consists of individual multiple-choice tests of Listening, Vocabulary, Reading and Grammar (with writing and speaking components under development). Similar to other tests based on this framework, reading, listening, and grammar items are text-based and aim to measure general and academic English proficiency, while the vocabulary items are discrete items and aim to measure student’s vocabulary knowledge. Despite the fact that a multiple-choice item format limits the items to assessing productive aspects of the English proficiency, this format was chosen to allow for immediate computerised marking.
The development and validation of DELTA

The Assessment lasts 90 minutes. Each component (except Vocabulary) consists of a number of parts. A DELTA assessment is constructed by the DELTA system to the structure shown in Table 1. The Listening and Reading Components consist of four parts and the Grammar Component consists of two parts. The DELTA system calculates the total number of items in these three components and then adds items to the Vocabulary Component such that the total number of items on the assessment equals 100.

Table 1
*The structure of the DELTA system*

<table>
<thead>
<tr>
<th>Component</th>
<th>Parts</th>
<th>Composition</th>
<th>Difficulty</th>
<th>Time allowed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Listening</strong></td>
<td>Part 1</td>
<td>1 Recording + 4–6 items</td>
<td>Easier</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Part 2</td>
<td>1 Recording + 6–8 items</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Part 3</td>
<td>1 Recording + 6–8 items</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Part 4</td>
<td>1 Recording + 6–8 items</td>
<td>More difficult</td>
<td>20–25 minutes</td>
</tr>
<tr>
<td><strong>Vocabulary</strong></td>
<td></td>
<td>20–25 Items</td>
<td>A range</td>
<td></td>
</tr>
<tr>
<td><strong>Reading</strong></td>
<td>Part 1</td>
<td>1 Text + 4–6 items</td>
<td>Easier</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Part 2</td>
<td>1 Text + 6–8 items</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Part 3</td>
<td>1 Text + 6–8 items</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Part 4</td>
<td>1 Text + 6–8 items</td>
<td>More difficult</td>
<td>65–70 minutes</td>
</tr>
<tr>
<td><strong>Grammar</strong></td>
<td>Part 1</td>
<td>1 Text + 10–15 items</td>
<td>A range</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Part 2</td>
<td>1 Text + 10–15 items</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For a diagnostic test to be effective, it should emphasise the identification of learners’ weaknesses (Alderson, 2005; Buck, 2001). That is, the test should be as comprehensive as possible and be able to provide detailed information about the subskills involved in accomplishing tasks or answering test items in Listening, Reading and Grammar. Thus, the DELTA development team thought it essential to explore the subskills involved in each component. This was done through a literature search of taxonomies of subskills for each component and consultation with a number of experienced language instructors and item writers about their perceptions of the subskills employed in different contexts. From this, a long list of subskills was compiled. The list was consolidated through moderations and discussions to a manageable number by eliminating obviously overlapping subskills and grouping similar subskills. Then the subskills were rewritten as *item intents* in accessible language, which the item writers, language advisors and students will easily understand in the DELTA student report. The following sections describe these processes for each component in more detail.
The DELTA Reading Component

The DELTA Reading Component tests students’ ability to read and understand the kinds of written English that they would read for English language learning and tertiary level study more generally. The texts that students read are of various text types (e.g., academic articles, news articles, and feature articles). These are the types of text that a Hong Kong tertiary student would be expected to be exposed to in an English language environment. Each text is written in standard English.

In order to provide useful diagnostic information for students, it is essential to decompose the construct of reading comprehension ability into a set of cognitive skills based on a substantive theory that describes the processes which a learner carries out when doing reading tasks (Jang, 2009). However, to date, no such theory has been developed. Over the last few decades, various lists of subskills have been drawn up by researchers; Alderson and Lukmani (1989) provide a very comprehensive summary of these lists. Davis (1944) conducted the first psychometric analysis and determined that there were nine comprehension skills which he later refined to eight (1968). Spearritt (1972) reanalysed Davis’ data and identified four separate factors: recalling word meanings; drawing inferences from the context; recognizing a writer’s purpose, attitude, tone and mood; and following the structure of a passage. However, he concluded that apart from vocabulary, which is the best differentiated, the three other skills were very highly correlated and thus could be predominantly measuring a single basic ability which he labeled as the ‘reasoning in reading’ (p.110).

Instead of devising questions to measure different subskills and then verify them empirically, Munby (1981) devised a list of ‘microskills’ (p. 19) that he considered to contribute to readers’ abilities to understand texts. However Alderson and Lukmani (1989) commented that Munby arrived at his list of skills not by devising questions to measure different subskills and then subjecting them to empirical verification, but by thinking about the nature of language and speculating on what a reader must logically have to do in order to understand a written text. Instead of using Munby’s microskills, Alderson and Lukmani (1989) used a different set of reading skills developed for the analysis of a reading comprehension test at the University of Bombay (Mumbai), which included recognition of words, identification, discrimination, analysis, interpretation, inference, synthesis and evaluation. More recently, Jang (2009) identified nine primary comprehension skills by analyzing think-aloud protocols with seven ESL students from a TOEFL preparation course and four graduate ESL students in the U.S.; and by recruiting five raters to evaluate the reading skills identified from the think-aloud protocols.

The construct incorporated into the DELTA Reading Component consists of subskills based mainly on the eight skills identified in the study by Alderson and Lukmani (1989) and with reference to Jang (2009); with modification and rephrasing by the DELTA team to make the intents as accessible to students as possible.

The DELTA Reading Subskills are shown below:

1. Identifying specific information
2. Interpreting a word or phrase as used by the writer
3. Understanding main ideas and supporting ideas
4. Understanding information and making an inference
5. Inferring the writer’s reasoning
6. Interpreting an attitude or intention of the writer
7. Understanding grammatical relationships of words or phrases across text
8. Identifying text type

In addition to being accessible to students, the DELTA Reading Subskills were written such that they are coherent with learning materials designed to help students to improve these subskills. In the DELTA Report that students receive after taking the Assessment, the test items that the students answered are displayed in terms of the subskills and are hyperlinked to relevant learning resources (see Figure 7 on p. 77). This is the case also for the Listening Component, which is described in the next section.

The DELTA Listening Component

The DELTA Listening Component aims to assess students’ ability to understand different types of spoken input in the university contexts. The recordings that students listen to are of various text types (e.g., debates, discussions, and conversations). These are the types of oral input that a Hong Kong tertiary student would be expected to be exposed to in an English language environment. Each recording is in standard English. A balance of English accents is used in each recording, i.e., Hong Kong, British, American, Australian, and Canadian, and there is a balance of female and male speakers. The speakers in the recordings use a natural rate of speech.

It is widely agreed that listening comprehension does involve a variety of subskills (Field, 1998; Field, 2005; Lund, 1990; Richards, 1983; Song, 2008) although Buck and Tatsuoka (1998) claimed that there was no consensus on precisely what the listening subskills are. In addition, Buck (2001) stated that the relatively small number of diagnostic listening tests available might have been partially due to the limited understanding of listening subskills. Richards (1983) suggested linking these subskills to existing listening proficiency descriptors, as most language proficiency descriptors are composed of a number of ‘can-do’ and ‘cannot-do’ statements; and comparing the subskills with the descriptors may help teachers identify which subskills students of a particular listening proficiency level need to focus on in their study. This argument echoes the idea that listeners use a number of subskills in comprehension and it is difficult to determine which subskill is the major one used in answering one particular question (Brindley, 1997). It is also argued that the skills tested in a listening test are not specific enough for diagnosis, which is a concern for Hughes (2002), who renders it worthwhile to include lower level listening skills such as discriminating between phonemes and prosodic features. This was the challenge facing the developers of the DELTA. Faced with similar constraints to the listening test in DIALANG (Alderson, 2005), the developers needed a test consisting of discrete items that would provide useful information to test users on the listening subskills that learners should work on so as to improve their overall listening proficiency.

To summarise, the skills or abilities usually tested in listening comprehension tests and in the DELTA Listening Component fall into three categories:
To discriminate sounds
To understand linguistic/literal meanings (specific information, main idea)
To understand inferred meanings/implications (speaker’s point of view, attitude)

Similar to the Reading Component, the DELTA Listening Subskills were written so as to be accessible to students and consistent with the ways that learning materials are designated:

1. Identifying specific information
2. Interpreting a word or phrase as used by the speaker
3. Understanding main ideas and supporting ideas
4. Understanding information and making an inference
5. Inferring the speaker’s reasoning
6. Interpreting an attitude or intention of the speaker

As can be observed, the DELTA Listening Subskills very closely resemble the Reading Subskills. This is deliberate so as to make it as straightforward as possible for student users to interpret the DELTA Report.

The DELTA Grammar Component

The DELTA Grammar Component tests students’ ability to identify and correct grammatical errors that might be made by Hong Kong tertiary student writers. The Grammar Component consists of two parts—a text and a number of multiple-choice test items. Each item tests a grammatical item (e.g., articles and verb tenses). Students read the text and select their answers to the test items. Items are identified as occurring in the text by an underlined section in the text. The student should choose from options A, B and C which consist of possible alternatives to the underlined section. Option D is always “no change” indicating that the original version is correct. The texts chosen originated from authentic texts written by tertiary level students in Hong Kong and are of varied text types (e.g., narrative, expository, and argumentative texts). They were modified so that any errors in parts of the text that are not underlined, i.e., are not tested, do not impede the student test-taker’s ability to answer the items.

The DELTA Grammar Component was based on a several definitions of the construct of grammatical knowledge. Rea-Dickens (1991, p. 114) defined grammar as ‘the single embodiment of syntax, semantics and pragmatics’, whereas Larsen-Freeman (1991) characterized grammatical knowledge along three dimensions: linguistic form (accurate), semantic meaning (meaningful), and pragmatic use (appropriate). According to Purpura (2004), grammatical knowledge embodies two highly related components: grammatical form and grammatical meaning. Grammatical form refers to linguistic forms on the subsentential, sentential and suprasentential levels. Grammatical form includes a host of forms, for example, on the phonological, lexical, morphosyntactic, cohesive, information management, and interactional levels. Knowledge of grammatical form, therefore, refers to the knowledge of one or more of these linguistic forms. Knowledge of grammatical meaning refers to knowledge of the meaning associated with an utterance as the sum of its parts and how these parts are arranged in syntax (literal meaning), as well as
how these parts are used to convey the speaker’s (or writer’s) intended meaning in context (intended meaning). Grammatical ability is the combination of grammatical knowledge and strategic competence; it is specifically defined as the capacity to realize grammatical knowledge accurately and meaningfully in testing or other language-use situations (Purpura, 2004).

Despite the variations in definition, there seems to be a consensus that correct formation of grammatical structures and accurate conveyance of meaning are key skills in demonstration of grammatical knowledge. Thus, in concert with Yin, Sims and Cothran (2012), a more traditional and conservative construct definition is adopted for the DELTA Grammar Component. In addition to the traditional categories of morphology and syntax used in the DIALANG test of grammar (Alderson, 2005), the DELTA Grammar Component also includes discursive (suprasentential) elements such as cohesive devices in its scope as the test uses students’ writing scripts as the raw material for item design. It is aimed to assess students’ grammatical ability in writing a coherent essay rather than using discrete grammatical points separately. Table 2 shows some of the grammatical elements or subskills used in the DELTA together with example items.

**Table 2**

*Examples of grammar subskills used in the DELTA*

<table>
<thead>
<tr>
<th>Grammar Subskills</th>
<th>Examples (Error is indicated by brackets and the correct answer by BLOCK letters.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjective</td>
<td>John is (friend) FRIENDLY to his new roommate.</td>
</tr>
<tr>
<td>Adverb</td>
<td>Thomas ran more (quick) QUICKLY than Jill.</td>
</tr>
<tr>
<td>Cohesive device (including conjunction and prepositional phrase used as cohesive device)</td>
<td>She was poor (and) BUT she was honest.</td>
</tr>
<tr>
<td>Determiner</td>
<td>She had a long holiday in (the Asia) ASIA last year.</td>
</tr>
<tr>
<td>Gerund</td>
<td>I look forward to (meet) MEETING you next week.</td>
</tr>
<tr>
<td>Infinitive</td>
<td>It is easy (making) TO MAKE mistakes if you are not careful.</td>
</tr>
<tr>
<td>Object pronoun</td>
<td>The new chief is related to (he) HIM.</td>
</tr>
<tr>
<td>Participle (present or past participle used as attributive or adverbial)</td>
<td>They say she died of a (broke) BROKEN heart.</td>
</tr>
<tr>
<td>Phrasal verb</td>
<td>I could tell she was angry as she (hung down) HUNG UP on me while we were talking on the phone.</td>
</tr>
<tr>
<td>Preposition</td>
<td>The personalities of the children are a very big concern (of) FOR parents.</td>
</tr>
<tr>
<td>Reflexive pronoun</td>
<td>I think they need to take up the responsibility (them) THEMSELVES.</td>
</tr>
<tr>
<td>Relative pronoun</td>
<td>The main idea for the project, (who) WHICH is led by Professor Chan, is to design a better system.</td>
</tr>
<tr>
<td>Singular/plural (phrases)</td>
<td>Children need this kind of support in their daily (life) LIVES.</td>
</tr>
<tr>
<td>Subject/Verb agreement</td>
<td>The police (is) ARE doing all they can to deal with the problem.</td>
</tr>
</tbody>
</table>
Multiple choice format is considered an effective method of item design for the Grammar Component ‘mainly due to its clear advantage of utilizing distractors to identify the difficulties experienced by learners’ (Koizumi et al., 2011, p. 6). Using the common learner mistakes that are seen in their performance is a recommended technique for creating effective distractors in multiple-choice test construction (Downing, 2006; Henning, 1987).

The DELTA Vocabulary Component

The DELTA Vocabulary Component aims to test students’ ability to read and understand the English words and phrases that they would encounter as a part of their English language learning and tertiary level study generally. In other words, the test aims to assess students’ knowledge of a specific range of vocabulary—their academic vocabulary knowledge. Vocabulary items are written to assess the ability to recognize the appropriate word for the given context. Students are given a sentence with the targeted vocabulary word as a gap-fill and given four options to choose from.

Read (2007) claims that a vocabulary assessment can reveal the extent of the lexical gap second language users face in coping with authentic reading materials and undertaking other communicative tasks in the target language. Thus, a vocabulary test can be seen as both a predictor and indicator of learners’ language proficiency. Because the DELTA is specifically designed for tertiary level students, Coxhead’s (2000) Academic Word List (AWL) was deemed the appropriate word list to be tested in the DELTA Vocabulary Component.

The AWL is a list constructed by frequently used words that are shared by texts of different types from various academic subjects. In spite of its limitations, it was chosen because it provides a useful basis for testing the general academic vocabulary level of Hong Kong university students. Although there are other word lists available, for example, the GSL (West & Jeffery, 1953) and the University Word List (UWL) by Xue and Nation (1984), they are not suitable for a number of reasons. The GSL contains words from more non-academic texts than academic texts, which is beyond the scope of the DELTA. The GSL was also compiled in the 1980s and might be unsuited to current academic learning needs (Nation, 2004). The UWL, even though it was created by a combination of four other word lists from various researchers (Campion & Elley, 1971; Ghadessy, 1979; Lynn, 1973; Praninskas, 1972), is considered too small and does not contain a large enough range of text types. Varying principles used to select words in each of the four word lists is another problematic issue of the UWL. Considering these issues, the AWL is comparably a better match of the DELTA’s needs. Moreover, the AWL provides readily available materials for both vocabulary learning and testing, such as, links to vocabulary exercises. This also helps the DELTA in informing both the teaching and learning of both students and language instructors.

Test Data Analysis and Score Reporting

Rasch measurement was used to calibrate items and generate diagnostic reports. The Rasch model (Rasch, 1960) is a prescriptive probabilistic item-response theory (IRT) model that aims to identify the capacity of a test to measure a latent variable (Bond & Fox, 2007; McNamara & Candlin, 1996). Through the Rasch model, one is
able to identify statistical patterns in performance by candidates (persons) on test items in order to draw conclusions about the underlying difficulty of items and the underlying ability of candidates (McNamara & Candlin, 1996). In other words, it allows for the measurement of persons and construct by converting raw data into one equal-interval scale. This conversion is achieved by measuring the probability of a person getting an item right. A strict set of rules, called fit statistics, checks for the quantity of the latent variable measured (i.e., logits), the precision of the measure (i.e., standard error) and the quality of this measure estimation (i.e., mean square and $t$ value). Table 3 shows the acceptable ranges of these fit statistics. If the fit statistics are within the acceptable range, then the data is said to ‘fit’ the model, and thus serve as evidence of test validity (Baghai, 2011; Bond & Fox, 2003; Smith Jr, 2001; Wolfe & Smith Jr, 2007).

Table 3

*Indicators of good fit statistics (Linacre, 2007)*

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logits</td>
<td>should have a range of logits (high to low)</td>
</tr>
<tr>
<td>Standard error</td>
<td>close to 0</td>
</tr>
<tr>
<td>Mean-square</td>
<td>between 0.7 – 1.3</td>
</tr>
<tr>
<td>$t$ value</td>
<td>between ±2</td>
</tr>
</tbody>
</table>

Test scores are exported from the DELTA system and imported into WINSTEPS v3.73 (Linacre, 2011), a Windows-based Rasch measurement software. Diagnosis of test validity and reliability starts with the investigation of fit statistics of the items. If there are items that do not fit the model, they are discarded from the analysis. Next, the spread and suitability of items is also checked against person ability. The item-person (see Figure 1) shows how the DELTA items cover a range of ability levels. The left side of the figure is the Rasch logit with the item mean set to zero (0). More able persons are at the top of the scale and less able students are at the bottom. On the right side of the scale are the items with more difficult items at the top of the scale and easier items at the bottom of the scale.
Note: Each "#" is 25 persons. Each "." is 1 to 24 persons. M = the mean of the person or item estimates. S = one (1) standard deviation from the mean. T = two (2) standard deviations from the mean.
WINSTEPS table 3.1 shows the summary statistics of 2011 DELTA test analysis (see Figure 2). The mean measure of the persons is 0.58. The least able student had an ability estimate of -1.83 and the most able student had an ability estimate of 3.22. Item difficulties ranged from 3.83 to -5.69. A further 176 items in the item bank are not included in the analysis as they were either not used for this iteration of the test or were misfitting.

<table>
<thead>
<tr>
<th>Item difficulties</th>
<th>3.83</th>
<th>2.0</th>
<th>1.37</th>
<th>3.83</th>
<th>1.57</th>
<th>-5.69</th>
<th>1.83</th>
<th>1.44</th>
<th>-1.83</th>
<th>3.22</th>
<th>2.0</th>
<th>1.37</th>
<th>3.83</th>
<th>1.57</th>
<th>-5.69</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAX.</td>
<td>85.0</td>
<td>99.0</td>
<td>3.22</td>
<td>1.44</td>
<td>1.83</td>
<td>3.22</td>
<td>1.44</td>
<td>1.83</td>
<td>3.22</td>
<td>3.22</td>
<td>2.0</td>
<td>1.37</td>
<td>3.83</td>
<td>1.57</td>
<td>-5.69</td>
</tr>
<tr>
<td>MIN.</td>
<td>1.0</td>
<td>2.0</td>
<td>-1.83</td>
<td>.23</td>
<td>.32</td>
<td>-3.2</td>
<td>.15</td>
<td>-2.4</td>
<td>.32</td>
<td>-3.2</td>
<td>2.0</td>
<td>-1.83</td>
<td>.23</td>
<td>.32</td>
<td>-3.2</td>
</tr>
</tbody>
</table>

Figure 2. WINSTEPS table 3.1 – Summary statistics
WINSTEPS table 27.1 shows the summary statistics of student performance to individual test components (see Figure 3). The ‘Mean Measure’ shows that on average, students found grammar easiest, followed by listening and vocabulary, and then reading as most appropriate to the average ability of the students.

<table>
<thead>
<tr>
<th>ITEM COUNT</th>
<th>MEAN</th>
<th>S.E.</th>
<th>OBSERVED</th>
<th>MEDIAN</th>
<th>MODEL MEASURE</th>
<th>MODEL SEPARATION</th>
<th>MODEL RELIABILITY</th>
<th>CODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>647</td>
<td>-0.01</td>
<td>0.05</td>
<td>1.37</td>
<td>-0.02</td>
<td>7.64</td>
<td>0.98</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>209</td>
<td>0.56</td>
<td>0.10</td>
<td>1.44</td>
<td>0.58</td>
<td>6.48</td>
<td>0.98</td>
<td>Gram</td>
<td></td>
</tr>
<tr>
<td>194</td>
<td>0.08</td>
<td>0.09</td>
<td>1.26</td>
<td>-0.08</td>
<td>7.86</td>
<td>0.98</td>
<td>List</td>
<td></td>
</tr>
<tr>
<td>205</td>
<td>0.45</td>
<td>0.08</td>
<td>1.18</td>
<td>0.46</td>
<td>7.40</td>
<td>0.98</td>
<td>Read</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>0.12</td>
<td>0.22</td>
<td>1.33</td>
<td>0.17</td>
<td>18.61</td>
<td>1.00</td>
<td>Voc</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. WINSTEPS table 27.1 in logit values

Another evidence of test validity is to determine the spread of item measures against person measures; do the items measure a range of abilities and is there a range of person abilities. These are derived from a calculation of the item/person strata. This statistic indicates the number of statistically distinct levels of item difficulty separated by at least three errors of measurement (Wright & Masters, 2002). Item/person strata is initially based on the item or person separation value (see Figure 2) and calculated using the formula in Wright and Masters (2002, p. 888): Item/person strata = ((4 x item/person separation) + 1)/3. Person strata is 2.70 and this indicates that the students who took the test can be grouped into three statistically distinct levels of proficiency. Item strata is 7.68 which indicates that the test takers can statistically distinguish at least eight levels of item difficulty.
The development and validation of DELTA

To avoid misconceptions about DELTA scores’ equivalence with percentages or raw scores, Bond and Fox (2007) suggest converting logit values to a scale with only positive values. Logit values are thus converted to a 200-point scale with 1 logit = 11.574 DELTA points (UMEAN=101.273; USCALE=11.574). Figure 4 shows the summary statistics in DELTA Measures.

<table>
<thead>
<tr>
<th>ITEM</th>
<th>MEAN</th>
<th>S.E. OBSERVED MEAN</th>
<th>S.D. OBSERVED MEDIAN</th>
<th>MODEL MEAN</th>
<th>S.E. MODEL</th>
<th>MODEL S.D.</th>
<th>SEPARATION</th>
<th>RELIABILITY</th>
<th>CODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>647</td>
<td>101</td>
<td>1</td>
<td>16</td>
<td>101</td>
<td>7.63</td>
<td>.98</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>209</td>
<td>95</td>
<td>1</td>
<td>17</td>
<td>95</td>
<td>6.46</td>
<td>.98</td>
<td>G</td>
<td></td>
<td></td>
</tr>
<tr>
<td>194</td>
<td>102</td>
<td>1</td>
<td>15</td>
<td>100</td>
<td>7.85</td>
<td>.98</td>
<td>L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>205</td>
<td>106</td>
<td>1</td>
<td>14</td>
<td>107</td>
<td>7.39</td>
<td>.98</td>
<td>R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>103</td>
<td>3</td>
<td>15</td>
<td>103</td>
<td>18.64</td>
<td>1.00</td>
<td>V</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SUBTOTAL RELIABILITY=.87
UMEAN=101.273 USCALE=11.574

Figure 4. WINSTEPS table 27.1 in DELTA values

At present, the DELTA measures are points on the DELTA proficiency scale. It allows one to track progress (or growth) each time a student takes the DELTA (see Figure 5). As more data is collected, the ‘meaning’ (e.g., weak vs. competent) or description of this measure against other proficiency tests (e.g., HKDSE) or frameworks (e.g., CEFR) will be established.

Figure 5. Sample DELTA track†
Finally, each student’s individual component measure is compared against the overall measure through a differential person function (DPF) analysis. The DPF analysis shows a student’s relative performance on each of the four test components of the test against the overall DELTA measure. The results of the DPF analysis are used to generate the component skills profile of the DELTA Report, which is described in the next section.

The DELTA Report

All items in the DELTA are machine-scored. Each item contains the following information:

1. Item difficulty
2. Language subskills
3. Text type (for Reading and Listening)
4. Text theme (for Reading and Listening)

All items are calibrated using live test data. The item difficulty of each item answered correctly contributes to the student’s overall proficiency or DELTA Measure (see Figure 6). The language subskills tested by each item are recorded in the student’s DELTA Report (see Figure 7). Statements of the student’s strengths and weaknesses in the language subskills are based on the item difficulty relative to the student’s proficiency (e.g., DELTA Measure 113 in this case). In other words, items that are of a lower difficulty level than the student’s proficiency are those that the student would be expected to answer correctly. If they are not answered correctly, they indicate possible weakness in that particular subskill. When students receive their reports, they are then advised to make use of it through independent learning or with the help of their teachers. Teachers have been briefed on how to help students understand the report and advise them on suitable actions to take.

![Component Skills Profile](image)

*Figure 6. Sample DELTA component skills profile*
The development and validation of DELTA

Figure 7. Sample DELTA component diagnostic report on reading

DELTA 2011 Test Results

Development work on the DELTA system was completed in May 2011. After small-scale pilot tests of the new system at each of the participating institutions, full-scale implementation began in September 2011. Since then, a total of 2947 students have taken the DELTA (see Table 4). There were 382 students whose results were not useful for the analysis and so only 2565 students were considered for calibration of the test items (see above). However, all students who took the DELTA received a DELTA Report on their performance.

Table 4
Numbers of students to have taken DELTA (September 2011–June 2012)

<table>
<thead>
<tr>
<th>Institution</th>
<th>No. of students</th>
</tr>
</thead>
<tbody>
<tr>
<td>City University of Hong Kong</td>
<td>2033</td>
</tr>
<tr>
<td>Lingnan University</td>
<td>488</td>
</tr>
<tr>
<td>Hong Kong Polytechnic University</td>
<td>426</td>
</tr>
<tr>
<td>TOTAL</td>
<td>2947</td>
</tr>
</tbody>
</table>

The majority of test takers have been at City University, where taking the DELTA each year is a requirement for all Year 1 students, as it is at Lingnan University, though the latter has a smaller intake. At the Polytechnic University, all the students...
who took the DELTA did it on a voluntary basis. The following paragraphs will describe the results of the DELTA testing in 2011–12.

Table 5 shows the descriptive statistics for the DELTA testing results. The mean DELTA Measure on each component was about 108, though the range of scores varied as indicated by the standard deviations, with the greatest range shown by Vocabulary and the lowest by Reading.

Table 5
Descriptive statistics DELTA 2011–12

<table>
<thead>
<tr>
<th>Component</th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>2525</td>
<td>66.0000</td>
<td>151.0000</td>
<td>108.346931</td>
<td>9.5694163</td>
</tr>
<tr>
<td>Listening</td>
<td>2564</td>
<td>77.0000</td>
<td>150.0000</td>
<td>108.113105</td>
<td>9.1652621</td>
</tr>
<tr>
<td>Reading</td>
<td>2526</td>
<td>74.0000</td>
<td>137.0000</td>
<td>107.995249</td>
<td>8.4043822</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>2526</td>
<td>68.0000</td>
<td>152.0000</td>
<td>108.739113</td>
<td>10.9196721</td>
</tr>
<tr>
<td>OVERALL</td>
<td>2565</td>
<td>80.0000</td>
<td>139.0000</td>
<td>107.991033</td>
<td>6.9339652</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>2524</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 8 shows the overall DPF analysis results in DELTA Measures for all students relative to the mean DELTA Measure. The students performed best at Vocabulary, followed by Grammar, Listening, and Reading.

Figure 8. Component skills profile of all students who took the DELTA in 2011–12

In order to find out if students from each university performed significantly differently on the four DELTA components, a one-way ANOVA test of three universities on four variables (i.e., the average DELTA Measure on individual components per university) was conducted. Figure 9 shows the mean performance of students on each component per university. The results indicate that Lingnan students performed significantly better in Vocabulary compared to the other test components (p<0.05). For the other universities there was no significant difference in performance on the four components.
The results from the DELTA testing at the three universities at which the DELTA was administered in 2011–12 showed little difference across the three universities other than the finding that students at Lingnan University seemed to do better at Vocabulary than at the other three components. This could be a reflection of that institution’s emphasis on liberal arts, though this would be speculation. The results of future administrations of DELTA at Lingnan will need to be monitored to see if this trend continues and if so, the possible reasons for this will be explored. The DELTA allows for this kind of research by providing detailed results of students from different institutions, and from different disciplines and faculties within the institutions on an ongoing basis. This can help inform curriculum design so that the language needs of students are met. As a follow-up to the full implementation of the DELTA, the test-takers’ feedback on the test and on the diagnostic report were sought and findings indicate that the students found the report quite helpful (Tsang, forthcoming).

**Future Challenges for the DELTA**

This paper has described the development and operationalisation of the Diagnostic English Language Tracking Assessment (DELTA) in Hong Kong, including the initial results from the first administration. The main function of the DELTA is to inform students about their English language proficiency and to help them to monitor their progress as they seek to improve this proficiency while they are at university. To this end it is vital that the DELTA is supported through providing DELTA-dedicated learning resources, such as teacher/advisor support and self-access learning provision, both physical and online. These resources are gradually being put
in place at the three institutions through initiatives such as the Inter-university Collaborative Online Self-Access (ICOSA) project, which is developing an online repository of self-access English language learning resources for Hong Kong university students.

Perhaps the greatest challenge of the DELTA is to be able to predict growth in English language learning in students so that they will know just what they need to do to get to where they wish to be in terms of proficiency at graduation. As more students take the DELTA repeatedly, the data gathered will facilitate the modeling of the typical growth patterns of students who start at a particular proficiency level, engage in particular language learning activities through both a formal curriculum and extra-curricular activities, and study in particular universities in particular programmes. It is anticipated that in this way, the DELTA can make a significant contribution to facilitating students learning of and through English while they are at university.

Notes
1. From the 2013–14 academic year, all students entering higher education as Hong Kong school leavers will take the Hong Kong Diploma of Secondary Education at the end of six years of secondary education. In 2012–13, a double cohort graduated from high school, one consisting of those students who had taken the HKDSE and the other of those who had taken the old system of Hong Kong Advanced Level Examinations.
2. CEPAS will be discontinued after the 2013–14 academic year.
3. For a description of how IRT is most suitable for use in the analysis of results of computer-based tests, see Fulcher (2000).
4. This DELTA Track is a simulation for a student who has taken the DELTA each year on a four-year degree programme.

References


