A Hybrid of a CBT- and a CAT-based New English Placement Test Online (NEPTON)

Salomi Papadima-Sophocleous
University of Nicosia (Intercollege), Cyprus

ABSTRACT
In recent years, many tertiary institutions have been changing their pen-and-paper English placement test practices into computer based ones. In the process of constructing the University of Nicosia (Intercollege) *New English Placement Test Online* (NEPTON), we discovered how to redesign our test to include the use of technology. The present article reports this experience in three parts. In the first section, we demonstrate how NEPTON, a hybrid of a computer-based test (CBT) and a computer adaptive test (CAT) model was influenced and shaped by theoretical issues such as CBT and CAT characteristics—their advantages and disadvantages—and by practical issues such as the particular context of the institution. This is followed by a detailed description of the test, highlighting its essential and innovative features. Lastly, we discuss and analyze different aspects of the test's first administration to establish how reliable and valid this test is as a placement instrument for first-year University of Nicosia (Intercollege) students.

KEYWORDS
New English Placement Test Online (NEPTON), Hybrid of a Computer-based Test (CBT) and a Computer Adaptive Test (CAT), English Placement Test (EPT), Foreign Language Placement Testing, Second Language (L2), Foreign Language (FL)

INTRODUCTION

Just as language testers had to come to terms with continuous new changes in testing methodologies, today's language testers are deeply caught in the web of technological advances that affect second language (L2) testing. They need to learn how to redesign their tests, integrate a variety of new technologies and applications, and apply technology in ways that meet their testing aims and the needs of their test takers.

In 2003, the University of Nicosia (Intercollege), one of the major private institutions of higher education in Cyprus felt it was time to respond to the new L2 testing realities. The large question was how to change its existing English placement pen-and-paper practices to online ones. In view of this task, two major subsequent questions were identified as central to this change.

1. Which computer-based language testing model is the most suitable for our English placement testing program?
2. Can a hybrid of a computer-based test (CBT) and a computer adaptive test (CAT) serve as an efficient and reliable online language placement testing tool for large groups of English learners?

Focusing on these two questions, this article documents the theoretical and practical issues influencing the test model concept, how this model was shaped by theoretical and particular context considerations, and how the implementation and evaluation process of the *New English Placement Test Online* (NEPTON) was affected by these factors.

COMPUTER-ASSISTED LANGUAGE TESTING (CALT)

L2 Testing and Computers

Computers have played a key role in language testing since 1935 (Fulcher, 2000), particularly since the expansion of personal computers in the late 1970s and 1980s (Godwin-Jones, 2001), and many institutions have incorporated technology in their language testing. However, test developers still have to decide which test type and design would fit their language-testing needs.

Current English Placement Test (EPT) Practices

As in many other tertiary institutions, the University of Nicosia (Intercollege) employs scores of tests, such as *Test of English as a Foreign Language* (TOEFL), *International English Language Testing System* (IELTS), and *General Certificate of Education* (GCE), to make placement decisions for incoming ESL students. These tests exist in both pen-and-paper and computer-based form. They primarily use multiple-choice questions with some using other types of activities such as sentence completion, synonym selection, and fill in the blank. These tests are of a generic type. Although there is criticism of such standardized tests (Alderson, 1987, Fulcher, 2003), it is widely recognized that they are still one practical way of measuring students' language competence.

Many universities and colleges develop their own placement tests. Most of these are mainly pen-and-paper tests, testing reading comprehension and grammar in mainly multiple-choice question form and in essay writing form. Gradually, however, many institutions are moving from these practices to computer-based ones: *Calis*, Duke University; *Dasher*, University of Iowa; *MaxAuthor*, University of Arizona (Godwin-Jones, 2001), *Quick Placement Test* (2001), Oxford and Cambridge Universities, and other agencies (*DIALANG*, 2001). Free or commercial electronic assessment devices are also used (e.g., *Question Mark*, 2004; *Hot Potatoes*, 2004; and *Quia*, 1998-2006). These devices offer a substantial variety of techniques used for electronic language testing. However, they are generic and restrictive in their functions because they only provide templates of specific test activity types, and, more important, they depend on central, external control and can prove quite expensive. Assessment electronic authoring tools, which form part of comprehensive online learning environments, such as *WebCT* (2004) and *Blackboard* (2004), are also used over the Internet (Godwin-Jones, 2001) for such needs. Since the University of Nicosia (Intercollege) uses *WebCT* to deliver courses, it was important to study its testing features. These features offer test formats such as multiple choice, essay, gap filling, and matching/ranking. However, they have the same restrictive characteristics as generic tests in that they lack the flexibility in areas where the developer of the authoring system has made decisions that may not harmonize with the wishes of the test developer (Chapelle & Douglas, 2006). Many electronic testing techniques derive from techniques suggested in the L2 testing theory literature for pen-and-paper tests (Hughes, 1989; Weir, 1990; Heaton, 1988). However, the interactivity offered by electronic testing devices enriches the pool of testing techniques. From this review, it was clear that our institution needed a test tailored to its students' needs, reflecting our syllabi, built and controlled by the institution, and not expensive to use.

Computer-based Tests (CBTs)

After establishing the extent of CBT practices, the question of which computer-based model would be most suitable for the University of Nicosia (Intercollege) English Placement Testing program was addressed. The study of CBT and CAT helped in deciding which computer-based model would fit our institution's needs. Computerized testing, whether adaptive or not, provides some significant advantages (Brown, 1997; Dunkel, 1999; Roever, 2001) over pen-and-paper testing: CBTs are less expensive (no paper waste, no markers), and they usually take less time and are therefore more efficient. Research has shown that a 50-question computerized test takes much less time to administer than a 50-question pen-and-paper test. Many test takers like computers and are therefore more motivated and even enjoy the testing process (Stevenson & Gross, 1991). Taking a test on the computer seems to be more interesting and less intimidating than taking it on paper. Test takers may find that CBTs are less overwhelming (as compared to equivalent pen-and-paper tests) because the questions are presented one at a time on the screen rather than in an intimidating test booklet with hundreds of test items. Electronic tests can also provide more variety in testing techniques, which can prove more interactive and authentic. New types of questions (e.g., point and click, drag and drop, and simulations) improve the test's ability to measure important skills. CBTs can provide improved test security, consequently test results are more meaningful. Test takers make fewer extraneous errors answering computerized test questions than they do when filling in the small circles on answer forms for pen-and-paper tests. Computerized tests are much more accurate for scoring selected-response test results than are pen-and-paper tests or oral exams and for reporting scores than human beings are. Tests can be scored immediately, providing instantaneous feedback for the examiners and the examinees. Computers can also provide statistical analysis of results. CBTs are superior in terms of reliability and validity (Dunkel, 1991). The increased use of computers in test development and delivery may reduce the testing costs in the future for the test developer, test publisher, test user, and test taker.

Limitations of CBTs

We were however concerned for some CBT features: questions in CBTs are the same for all test takers, questions are presented in a linear form, CBTs usually have many questions and are rather long, and maximum marks give the score. These were aspects we did not want to include in our test.

Computer Adaptive Tests (CATs)

CATs have some additional advantages (Dunkel, 1991): they are more efficient than CBTs in that they are shorter and quicker because they use even fewer test items to arrive at the test taker's level. Even with their shorter time limits, CATs provide more time per test question compared to both pen-and-paper and CBTs. Moreover, CATs allow easier test revisions. If a test question is not functioning well, it can be removed without a complete republishing of the test. CATs also improve security in several ways. First, CATs expose items at a much reduced rate (large item pools, random choice, adaptive testing), allowing the items to be effective for a longer period of time. Second, test-coaching efforts that focus on individual items are less effective because it is not clear which items will be presented to the person. All these features make it impossible for one examinee to successfully cheat by copying another's correct answers.

According to theorists, the main advantage of CATs over traditional computerized test designs is efficiency: this is achieved by avoiding presenting questions that provide no help in determining the person's score (i.e., questions that are too easy or too hard). CATs are tailored to the needs of individual test takers (Chapelle & Douglas, 2006). Each test taker answers questions that are personally challenging without being too hard or too easy. Boredom from answering many easy questions and frustration from answering too many hard questions are avoided. This efficiency is the main reason, according to Grist (1989), that certification candidates overwhelmingly prefer adaptive tests and why many institutions have adopted this measurement technology. Because of this, it is unclear exactly when the test will end. A CAT usually presents a variable number of questions, and a minimum and maximum number of questions are typically set. The score is not based on the number of correct answers but is derived from the level of difficulty of the questions answered correctly. The score is computed statistically and is based on the principles of item response theory (Lord, 1980). Thus, CATs produce a similar psychological test-taking experience for everyone.

Limitations of CATs

Although CATs seem to tailor testing to the individual test taker's ability, there are some features we were concerned about. As Chappelle and Douglas (2006) argue, questions have been raised concerning the effect of leaving item selection up to a computer program that chooses items on the basis of the level of difficulty of the items. We agree with them that content should be taken into consideration, that items should not only be selected from the pool for any given candidate on the basis of statistical characteristics alone without making sure that the candidate appropriately sampled items from the relevant content, and that this should not be left to chance. Selection of items included on a CAT by an algorithm may not result in an appropriate sample of test content (Chapelle, 2001). Two approaches attempted to avoid having the item selection algorithm choose each item individually and gain more control over the way content is sampled: (a) the presentation of items in "testlets" (Wainer & Eignor, 2000)—in other words the clustering of items of different aspects of the construct of interest of the same level (e.g., different grammar aspects) presented in a passage—and (b) the tagging of individual items in a pool with information about content and having the item selection algorithm choose the items on the basis of content as well as statistical properties. We found that the first approach, although with some potential for further investigation, restricts itself to one aspect of language (e.g., grammar) and that the second approach is restricted to individual words or sentences. Another concern with this type of adaptive selection is the risk of having the test prematurely terminate if the calculated ability estimate triggers a search for a question with a difficulty level beyond the upper/lower limits of the item bank. We also felt concern for the fact that test takers are not given the same and adequate number of questions, are not tested adequately in the various skill areas and in various types of activities and text types, and are not tested in all levels (very often students know different aspects at different levels). Another concern is that CATs do not allow test takers to revisit test items. Test takers indicate greater satisfaction when they have the opportunity to review their answers and often tend to improve their answers if they have the opportunity to review them (Vispoel, Hendrickson, & Bleiler, 2000).

Disadvantages of CBT and CAT

Research also refers to some disadvantages using computers in L2 testing. Brown (1997), for example, divides these disadvantages into two categories: physical considerations (computer equipment not always available or in working order, reliable sources of electricity, screen size limitations, and graphics capabilities) and performance considerations (student familiarity with using computers or keyboards). Dunkel (1991) also talks about the high cost of computer hardware and software, the time required to acquaint test takers with the computer, and the potential of evaluating not only language but also computer skills. Cohen (1984) and Larson and Madsen (1985) are also concerned with the potential bias in computerized exams against students unfamiliar with the new technology. These disadvantages do not apply in our institution. There is a large number of computer labs continuously serviced by lab assistants. Computer equipment is always in working order, therefore delivering a test electronically using personal computers is not a problem. Our online tutorial familiarizes test takers with the test features and accommodates them with a test trial.

TEST CONSTRUCTION

We then investigated various theories in test construction and evaluation. Alderson, Clapham, and Wall (1995) discuss the importance of test specifications (purpose, learner, sections/papers, target language situation, text types, language skills, language elements, tasks, items, methods, rubrics, and criteria). Chapelle and Douglas (2006) suggest a four-process architecture for an assessment system (adapted from Almond Steingerg and Minlevy, 2002): activity selection process, presentation process, response process, summary scoring process. Fulcher (2003) considers various aspects of the process in designing an interface for a computer-based language test. He discusses general design considerations and processes: prototype (hardware and software), interface (navigation, terminology, page layout, text and text color, toolbars and controls, icons and graphics, help facilities outside the test, item types, multimedia, forms for writing and short-answer tasks, and feedback) and concurrent design issues (activities, usability testing, field testing, and fine tuning). Noijons (1994) suggests some points which can be used to evaluate two major CALT aspects: (a) aspects of test content—before taking the test (purpose and objective of the test, test length and ways items are selected from an item bank), during (item type and format of responses, feedback, test time, responses registration), and after (evaluation and test results presentation)—and (b) aspects of taking the test—before (entrance to the test, test instructions, examples of items, test check), during (fraud, breakdowns, feedback, end of test), and after (storage and printing of data). We took all of these into consideration and tailored many of them to the development of our own test.

BACKGROUND INFORMATION

After studying and critically analyzing the existing L2 testing models, we reviewed the existing University of Nicosia (Intercollege) EPT practices to establish the needs. Students at the University of Nicosia (Intercollege) come from government or public schools in Cyprus or from schools from approximately 78 other countries. They are evaluated upon their entrance and placed into six English language levels and respective courses (BENG-50, BENG-80, BENG-90, BENG-100, ENG-100 and ENG-101). Up until 2003, a pen-and-paper EPT was used to place students (Intercollege *Placement Test: Marking Guidelines and Sample Placement;* Intercollege *English Placement Test,* pen-and-paper format). This test had an item bank of 64 questions used in four different exam papers in a different item order. It covered structure (30 points), vocabulary through sentence-based multiple-choice questions (15 points), reading comprehension through text-based multiple-choice questions (19 points), and writing through essay writing. Students were placed based on the following system: Test score 0-18: BENG-50; 19-27: BENG-80; 28-36: BENG 90; 37-45: BENG 100; 46-54: ENG-100, and 55-above: ENG-101. For the essay the following formula was followed: essay score 0: BENG-50; 1: BENG-80; 2: BENG-90; 3: BENG-100; 4: ENG-100, and 5: ENG-101. When test score conflicted with essay score, the essay score took precedence. Various standardized exams such as GCE and TOEFL were also used. University of Nicosia (Intercollege) EPT was based on traditional pen-and-paper testing methods, administration, correction, and result reporting.

THE NEW ENGLISH PLACEMENT TEST ONLINE (NEPTON)

Test Specifications

After researching the existing computer based language testing models and University of Nicosia (Intercollege) EPT practices, we decided to develop our own test design for our *New English Placement Test Online* (the NEPTON test). The test specifications were designed based on Alderson et al.'s (1995) test specifications, Chapelle and Douglas (2006) assessment system architecture, Fulcher's (2003) interface design, and on Noijons's (1994) evaluation guidelines.

The NEPTON Purpose

NEPTON was developed to assess online University of Nicosia (Intercollege) incoming students' English language proficiency and place them in English courses appropriate to their level of English competence.

NEPTON Content

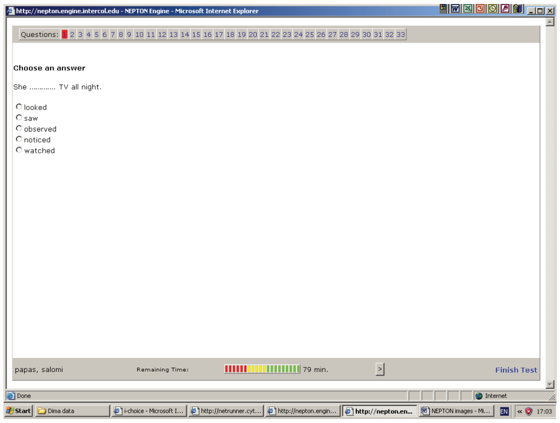
The comparison of current theories and practices in L2 testing to the particular University of Nicosia (Intercollege) testing context resulted in the adoption of a form of the 'loosely' used communicative paradigm as the design approach for NEPTON. This was based on the following facts: NEPTON had to serve the specific context of the University of Nicosia (Intercollege) and the particular needs of its students. A communicative language test is based on a description of the language that test takers need to use and reflect communicative situations in which test takers are likely to find themselves. For University of Nicosia (Intercollege) students, these situations are

1. Everyday situations
   Although the University of Nicosia (Intercollege) is in Cyprus, where Greek is the prominent language, English is mainly the common language of communication among non-Greek speakers who live or study in Cyprus. In other words, English is very often the language of communication in students' everyday life.

2. Academic setting
   Although in a predominantly Greek-speaking country, the University of Nicosia (Intercollege) uses English as the language of instruction, therefore students need an adequate knowledge of English for their academic needs.

According to the relevant literature, communicative language tests are those which make an effort to test language in a way that reflects how language is used in real communication (Kitao & Kitao, 1996). Although it is not always possible to make language tests fully communicative, it may often be possible to incorporate communicative elements in them. Following this approach in NEPTON, we included authentic or authentic-like text types, similar to the ones students would come across in their academic, personal, and social settings in Cyprus and overseas from newspapers, magazines, advertisements, literary books, letters, and short stories. These text types cover relevant topics, and incorporate vocabulary, structure, and sociolinguistic elements which derive from the literature review and the University of Nicosia (Intercollege) English curriculum. The test was designed to stress variety by including a sampling of different topics, functions, situations, levels of difficulty, lengths of passages, and types of questions. Contextual clues are strongly evident, so that items are functionally and semantically explicit. The test assesses writing using two different approaches. First, writing is divided into discrete levels, vocabulary, and grammar, and these elements are tested separately by the use of objective electronic testing activities. They test knowledge of vocabulary and grammar in the form of sentence-based multiple-choice items (see Figure 1).

Figure 1
Sentence-based Multiple-choice Items


http://nepton.engine.intercol.edu - NEPTON Engine - Microsoft Internet Explorer
Questions: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33
papas, salomi
Remaining Time:
79 min.
>
Finish Test
Done
Internet
Start
Dima data
i-choice - Microsoft I...
http://netrunner.cyt...
http://nepton.engin...
http://nepton.en...
NEPTON images - Mi...
EN
17:03

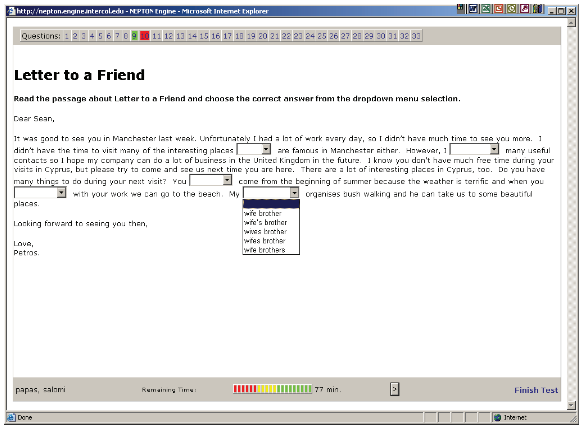
Choose an answer

She .... TV all night.

○ looked
○ saw
○ observed
○ noticed
○ watched

The test also assesses vocabulary and grammar knowledge in a more contextualized mode in the form of text with four or five multiple-choice questions of drop-down menu selections (see Figure 2).

Figure 2
Multiple-choice Question with Drop-down Menu Selections


http://nepton.engine.intercol.edu - NEPTON Engine - Microsoft Internet Explorer
Questions: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33
papas, salomi
Remaining Time:
77 min.
>
Finish Test
Done
Internet

Letter to a Friend

Read the passage about Letter to a Friend and choose the correct answer from the dropdown menu selection.

Dear Sean,

It was good to see you in Manchester last week. Unfortunately I had a lot of work every day, so I didn't have much time to see you more. I didn't have the time to visit many of the interesting places ▾ are famous in Manchester either. However, I ▾ many useful contacts so I hope my company can do a lot of business in the United Kingdom in the future. I know you don't have much free time during your visits in Cyprus, but please try to come and see us next time you are here. There are a lot of interesting places in Cyprus, too. Do you have many things to do during your next visit? You ▾ come from the beginning of summer because the weather is terrific and when you ▾ with your work we can go to the beach. My ▾ organises bush walking and he can take us to some beautiful places.

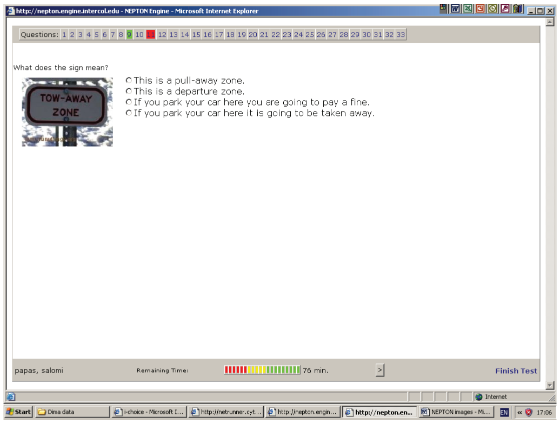
wife brother
wife's brother
wives brother
wifes brother
wife brothers

Looking forward to seeing you then,

Love,
Petros.

Reading comprehension is tested in two forms: in the contextualized and situational form of signs (accompanied by a visual) and texts with multiple-choice questions (see Figures 3 and 4).

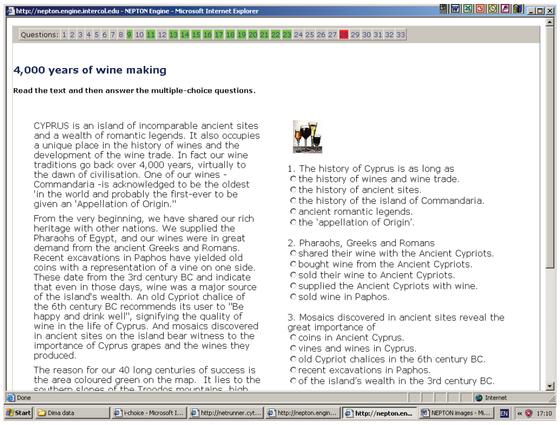
Figure 3
Contextualized Form with Multiple-choice Questions


http://nepton.engine.intercol.edu - NEPTON Engine - Microsoft Internet Explorer
Questions: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33
TOW-AWAY ZONE
papas, salomi
Remaining Time:
76 min.
>
Finish Test
Internet
Start
Dima data
i-choice - Microsoft I...
http://netrunner.cyt...
http://nepton.engin...
http://nepton.en...
NEPTON images - Mi...
EN
17:06

What does the sign mean?

○ This is a pull-away zone.
○ This is a departure zone.
○ If you park your car here you are going to pay a fine.
○ If you park your car here it is going to be taken away.

Figure 4
Text with Multiple-choice Questions


http://nepton.engine.intercol.edu - NEPTON Engine - Microsoft Internet Explorer
Questions: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33
Done
Internet
Start
Dima data
i-choice - Microsoft I...
http://netrunner.cyt...
http://nepton.engin...
http://nepton.en...
NEPTON images - Mi...
EN
17:10

4,000 years of wine making

Read the text and then answer the multiple-choice questions.

CYPRUS is an island of incomparable ancient sites and a wealth of romantic legends. It also occupies a unique place in the history of wines and the development of the wine trade. In fact our wine traditions go back over 4,000 years, virtually to the dawn of civilisation. One of our wines - Commandaria -is acknowledged to be the oldest 'in the world and probably the first-ever to be given an 'Appellation of Origin."

From the very beginning, we have shared our rich heritage with other nations. We supplied the Pharaohs of Egypt, and our wines were in great demand from the ancient Greeks and Romans. Recent excavations in Paphos have yielded old coins with a representation of a vine on one side. These date from the 3rd century BC and indicate that even in those days, wine was a major source of the island's wealth. An old Cypriot chalice of the 6th century BC recommends its user to "Be happy and drink well", signifying the quality of wine in the life of Cyprus. And mosaics discovered in ancient sites on the island bear witness to the importance of Cyprus grapes and the wines they produced.

The reason for our 40 long centuries of success is the area coloured green on the map. It lies to the southern slopes of the Troodos mountains, high

1. The history of Cyprus is as long as
○ the history of wines and wine trade.
○ the history of ancient sites.
○ the history of the island of Commandaria.
○ ancient romantic legends.
○ the 'appellation of Origin'.

2. Pharaohs, Greeks and Romans
○ shared their wine with the Ancient Cypriots.
○ bought wine from the Ancient Cypriots.
○ sold their wine to Ancient Cypriots.
○ supplied the Ancient Cypriots with wine.
○ sold wine in Paphos.

3. Mosaics discovered in ancient sites reveal the great importance of
○ coins in Ancient Cyprus.
○ vines and wines in Cyprus.
○ old Cypriot chalices in the 6th century BC.
○ recent excavations in Paphos.
○ of the island's wealth in the 3rd century BC.

Each item was selected according to the following criteria: the test purpose, topics, skills, format, sociocultural context and test takers' background and study setting, as described in the test specifications. Finally, the test assesses writing through a more direct, communicative, and extended global integrative writing task, in a nonelectronic hand-written mode.

In terms of content, this tailor-made test reflects the learning context, the students' profile, the type of English they use, their studies at the University of Nicosia (Intercollege), the English program at the University of Nicosia (Intercollege), and the social and educational context in Cyprus where the test is taken. The test is, to some extent, based on the old English placement practices (types of questions and language skill areas tested), so that the new test is smoothly accepted by stakeholders, regardless of the fact that it is also informed by current theories in L2 online testing.

NEPTON: A Hybrid of CBT and CAT

Taking the advantages and disadvantages of both CBT and CAT into consideration, we decided to explore the possibility of combining advantages from both tests and design a hybrid CBT/CAT which would attempt to avoid aspects of the tests we had concerns about.

Main Essential and Innovative Features of NEPTON

NEPTON is a computer-based test delivered online. It has unique and innovative features which make it different from CBTs and CATs. In NEPTON, items are not selected from a pool for any given test taker on the basis of the level of difficulty of the items and of statistical characteristics alone, but rather on an algorithm which systematically selects items also on the basis of their content. This algorithm selects items during the test from an item pool composed of subpools of substantial numbers of items of different language competence levels, different skills, and activity types. This process is based on a balanced and sufficient predetermined test choice and item selection of six level-based slides with nine items of a different skill and activity type per slide. This ensures that all test takers are tested at all levels through (a) a randomized selection of tests which offer a different predetermined selection of nine-item-slide-based selection and (b) a randomized selection from subpools of items of all skills tested, all test activity types, and in the same number of questions. The test is long enough to test all test takers adequately and short enough for test takers not to be bored. It offers similar psychological experience to all test takers. The nine-item slide selection algorithm, the large item pool, and the multiple randomization system also strengthen test security. Test takers are aware of the test length and the time available to take the test. Although items are administered one at a time and test takers are not overwhelmed with too many items presented together, they have at the same time the flexibility to browse through the items in any order they like at their own pace and review and change their responses. Both the test taker and the administrator interfaces have innovative features which make navigation and test use simple and friendly. The software, based on existing resources and local expertise, was made in house for more control over desired options. The database is stored according to the test specifications designed for the specific needs of the institution's test takers (choice of activity types, skills tested, item selection algorithm and the resulting cut off point system, and compatibility to the institution's placement levels). The hybrid nature of the test led to two more innovative features: a somewhat different and alternative system of item analysis and a system to calculate cut off points. As a result, each test taker is presented with a unique well balanced test of substantial length with skills tested and activity types at all levels.

Presentation Platform

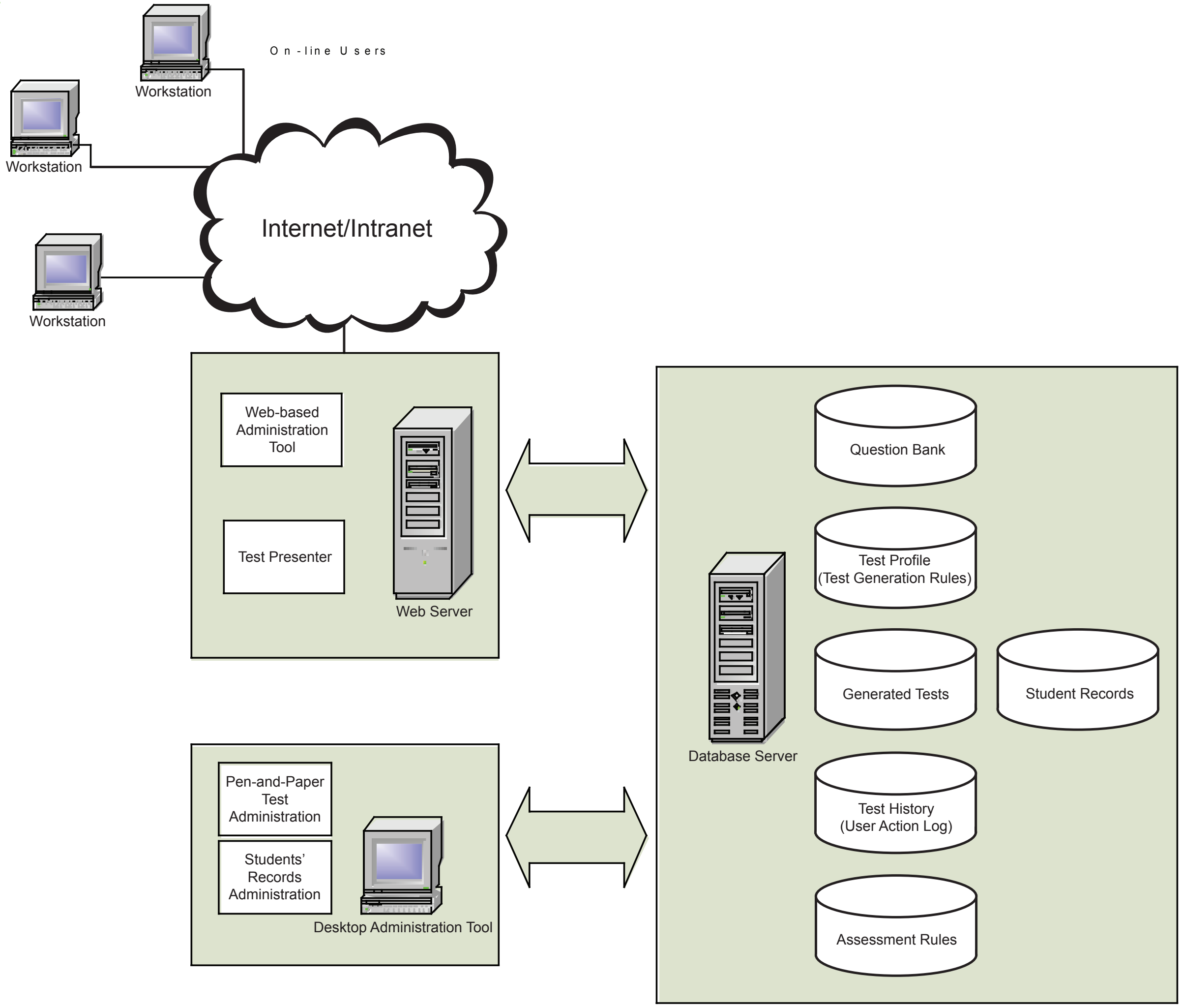
In the development of NEPTON, we had to consider computer hardware and software. Preference was given to the IBM/PC environment which was widely used in our computer labs. It was also decided to keep system requirements to a minimum: minimal RAM and graphics. We also kept the test programming structure modular and open ended so as to easily accommodate database changes.

The NEPTON Testing Software

The software was designed based on the test specifications, which included the features needed for the hybrid test. The system was developed on Microsoft.NET platform (Mack & Seven, 2002; Walther, 2003; Sceppa, 2002) and has the architecture shown in Figure 5.

The Test Database Server includes the Question Bank, the Test Profile, the Generated tests, the Students' Records, the Test History and the Assessment Rules. In the web server reside the Web Based Administrator and the Test Presenter. The Web Based Administrator provides the Question Database Management, the Test Profile Management, the User Management (Administrators), the Student Management (test takers), and the Reports Generating. The Desktop Administrator's tool includes the item editor, slide manager, user manager, and student manager, pen-and-paper test set up, system set up, change password, results, test key, and logout.

Figure 5
The NEPTON Test System Architecture

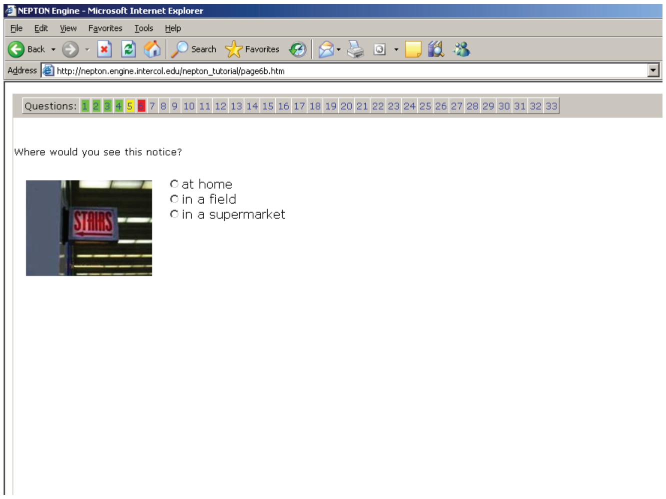

On-line Users
Workstation
Workstation
Workstation
Internet/Intranet
Web-based Administration Tool
Test Presenter
Web Server
Pen-and-Paper Test Administration
Students' Records Administration
Desktop Administration Tool
Database Server
Question Bank
Test Profile (Test Generation Rules)
Generated Tests
Student Records
Test History (User Action Log)
Assessment Rules

Test Taker Interface

The primary requirement of the test taker interface was that it be simple and user friendly. A tutorial gives test takers a description and samples of the four types of activities in their multiple-choice format, what those activities assess, and explains the various test interface and navigation features. The test takers have the opportunity to trial the test. After completing a short set of personal information input fields, they then take the test.

The test taker interface is very user friendly. The number of questions to be answered is clearly indicated at the top of the screen. The different button colors indicate the status of each item: green indicates questions already answered, yellow indicates that not all items in a text-based drop-down menu selection or multiple choice questions are completed and that the test taker needs to return and complete them, red indicates the current item, and grey indicates items still remaining to be answered (see Figure 6).

Figure 6
Test Taker Interface


NEPTON Engine - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://nepton.engine.intercol.edu/nepton_tutorial/page6b.htm
Questions: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33
STAIRS

Where would you see this notice?

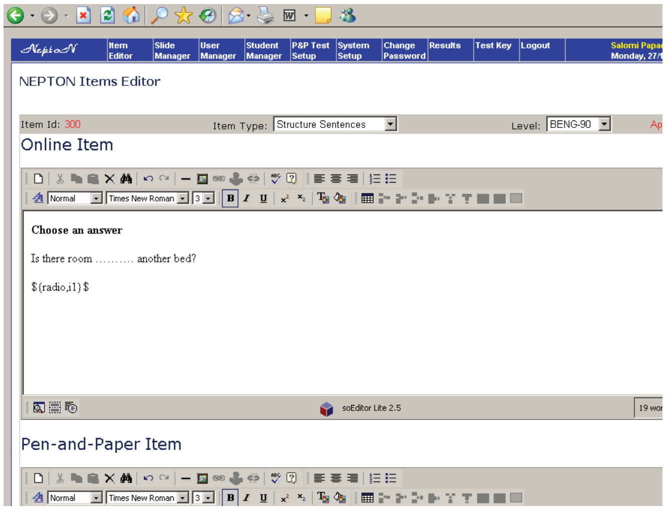
○ at home
○ in a field
○ in a supermarket

Test takers can move freely to any item in any order they wish. Each item is presented in the main area of the interface. At the bottom of the screen, test takers can see their name on the left, the time available and the *next-question* button in the middle, and the *Finish Test* button on the right. The only computer skills needed are mouse clicking and, for some text-based questions, scrolling.

The Administrator Interface

The Administrator interface is also user friendly. The menu bar at the top of the screen indicates the various functions: Item Editor, Slide Manager, User Manager, Student Manager, P&P test setup, system set up, change password, results, test key, and log out. At the top of the Items Editor area, the item identification number, type, and level are indicated. Here, the test administrator can upload test items according to their classifiers. In the upper (Online Item) section, the item is uploaded for online use (see Figure 7).

Figure 7
NEPTON Items Editor



NEPTON Items Editor

Item Id: 300 Item Type: Structure Sentences Level: BENG-90 Ap

Online Item

Choose an answer

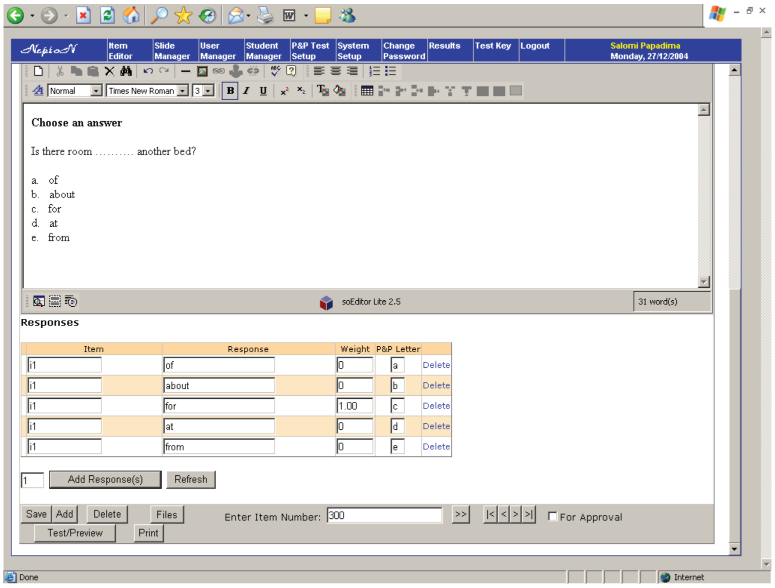
Is there room .... another bed?

${radio,i1}$

Pen-and-Paper Item

In the lower (Pen & Paper Item) section, the item is uploaded for the pen-and-paper version use (see Figure 8).

Figure 8
NEPTON Item Editor/Pen-and-paper Item Entries



Choose an answer

Is there room .... another bed?

a. of
b. about
c. for
d. at
e. from

Responses

Item	Response	Weight	P&P Letter	
i1	of	0	a	Delete
i1	about	0	b	Delete
i1	for	1.00	c	Delete
i1	at	0	d	Delete
i1	from	0	e	Delete

1

Enter Item Number: 300

□ For Approval

At the bottom of the page, the test administrator can edit or delete test items, include the responses of each multiple choice, their weight, and the letter for each one. Finally, the question can be previewed and then approved to become available. The navigation buttons are very easy to use. The rest of the functions of the upper toolbar of the administrator's interface are equally easy to use.

The Item Bank

We developed a large pool for the six English language competence levels. The items measure writing (structure and vocabulary) and reading comprehension skills (scanning and skimming). These discrete items represent a wide range from the whole area of content of the following activity types and assess the following skills: sentence-based structure (SB-S), text-based structure (TB-S), sentence-based vocabulary (SB-V), text-based vocabulary (TB-V), sign-based reading Comprehension (SB-RC), and text-based reading comprehension (TB-RC). These test items are stored in respective subpools.

MODERATION AND ITEM POOLS

The validity of the test items was examined through a moderation process. About 42% of the English program teaching faculty and four native speaker professionals in the field contributed to this process which involved: (a) editing, commenting on the clarity of each task, appropriateness of each item in terms of the level, appropriateness of what each item tested (vocabulary and structure) and (b) reading comprehension suitability of texts at each level, format and items, uniqueness of each correct answer, content, context and topics, and appropriateness of sociocultural aspects the embedded content. All these items were reviewed to qualify according to the criteria above. As a result, items were edited, choice or length of the texts was improved, and some deletions were made. The original number of questions and items was reduced to 632 questions consisting of 1,084 items. These were uploaded in the test software according to their classifiers (see Table 1).

Table 1
Test Questions and Test Items

Question

Level	SB-S	TB-S	SB-V	TB-V	SB-RC	TB-RC	Total
BENG-50	50	10	22	7	21	10	120
BENG-80	59	8	29	14	20	10	140
BENG-90	40	11	29	12	0	10	102
BENG-100	52	5	13	5	0	5	80
ENGL-100	46	5	55	6	0	4	116
ENGL-101	33	5	27	5	0	4	74
Total	280	44	175	49	41	43	632

Items

Level	SB-S	TB-S	SB-V	TB-V	SB-RC	TB-RC	Total
BENG-50	50	50	22	28	21	40	211
BENG-80	59	40	29	56	20	40	244
BENG-90	40	55	29	48	0	40	212
BENG-100	52	25	13	20	0	20	130
ENGL-100	46	25	55	24	0	16	166
ENGL-101	33	25	27	20	0	16	121
Total	280	220	175	196	41	172	1,084

QUESTION SELECTION ALGORITHM

The item selection algorithm was based on multiple randomizations. The test was generated through electronic and random choice following a nine-item slide design per level. There were two combinations of random selection: one for Test 1 and Test 2 and the other for item selection from the subpools for each item slide. Thus, if Test 1 is selected, three one-item questions are chosen from the SB-S BENG-50 50-item subpool, then one text (including 5 test items) from the TB-S BENG-50 10-item subpool and one question (one item) from the SB-RC BENG-50 21-item subpool (total of nine test items) (see Table 1). The random selection continues from the BENG-80 item pools to ENGL-101: (nine items per level) (see Test 1 in Table 2).

The same randomization process is followed if Test 2 is selected (See Test 2 in Table 2). The total length of each randomly generated unique test consists of six level-based slides of nine items each of 33 discrete questions equaling 54 items per student. This item selection ensures that all test takers take a unique test. They have the same number of questions and are tested at all levels. The test items are randomly generated for each level from different subpools covering different skills, text types, and content.

Table 2
Final NEPTON Slide Algorithm (After Field Testing)

	Test 1	Test 2
BENG-50		
9 items per slide	SB-S: 3 TB-S: 5 SB-RC: 1	VB-S: 4 VT or RC: 4 SB-RC: 1
BENG-80		
9 items per slide	VB-S: 4 VB-T / TB-RC: 4 SB-RC: 1	SB-S: 3 TB-S: 5 SB-RC: 1
BENG-90		
9 items per slide	SB-S: 4 SB-T: 5	SB-V: 5 TB-V or TB-RC: 4
BENG-100		
9 items per slide	SB-V: 5 TB-V or TB-RC: 4	SB-S: 4 TB-S: 5
ENG-100		
9 items per slide	SB-S: 4 TB-S: 5	SB-V: 5 TV-V or TB-RC: 4
ENG-101		
9 items per slide	SB-V: 5 TB-V or TB-RC: 4	SB-S: 4 TB-S: 5
9 items x 6 slides	54 items (33 questions)	54 items (33 questions)

THE NEPTON PEN-AND-PAPER FIELD TESTING

NEPTON was initially field tested in pen-and-paper form during the second semester of the 2003-2004 academic year. A sample of about 1,400 students participated in the field testing at three testing sites. Each test consisted of approximately 65 randomly chosen questions representing approximately 108 test items (two slides of nine items per level equalling 18 items by six level-based slides equalling a total of 108 items, and text-based questions included four to five test items each). The mode of item presentation followed a slide design from lower to higher levels, and each item was randomly chosen from the different test-item-type pools (sentence, cloze-text, signs, text, matching text, and titles). The aim was twofold: to have all test items used by as many students as possible so that item analysis could be as accurate as possible and to analyze the data in order to establish cut off points for use in the online trial of NEPTON.

ITEM ANALYSIS

The item analysis was done using the tests of 584 test takers and the 1,084 test items included in the test battery.

The NEPTON Facility Value (FV)

Based on the literature on item analysis (Alderson et al., 1955; Brown, 2003) and our specific research context, we decided that 25% to 80% would be an acceptable facility value (FV) for each item. To establish this, we had to come up with an alternative way of analyzing the data due to the test's hybrid nature: each item was ranked from high level to low level. The total number of test takers who took each item was recorded next to each item, together with the number of correct answers per test item. The total number of correct answers was then divided by the total number of test takers who took that item to establish the item's FV. Table 3 provides a sample of the data used for the item analysis. It shows how the test item data have been recorded and the way the FV of each item was established.

Table 3
FV: Sample Item Analysis

Item level	Item type	Item identity	Total no. of students who took the item (S)	Total no. of Correct Answers (C)	Facility Value (C/S)	Items with good facility value: 25%-80% (√)
ENGL-101	Reading comprehension	767/i4	14	12	0.85	
ENGL-101	Reading comprehension	767/i2	13	11	0.84	
ENGL-101	Vocabulary text	640/i2	34	27	0.79	√
ENGL-101	Reading comprehension	767/i3	14	11	0.78	√
ENGL-101	Structure text	556/i4	23	18	0.78	√
ENGL-101	Structure sentence	548/i1	22	15	0.68	√
Total of test items with good Facility Value						665/1,084

The NEPTON Discrimination Index (DI)

There are no rules as to what discrimination indices (DIs) are acceptable because the possibility of getting high DIs varies according to the test type and range of ability of the examinees. According to Alderson et al. (1995), the highest discrimination possible is +1.00, although item writers are often content with DIs of +.4 or above. In the *Statistics Guide: Interpreting the Reports* (n. d.), it is argued that "Ideally, test items will have a positive discrimination index above 0.30 … ." We used this figure to calculate the discrimination index of each test item of NEPTON. To do so, all 1,084 items were ranked from high to low level. The total number of test takers who took each item was recorded next to each item in two categories: high score and low score groups. The number of correct answers in the high score group (CH) and the total number of test takers who took each item at high score group (SH) were recorded next to each item. The difficulty index was then calculated for the high score group (DH). In addition, the number of correct answers in the low score group (CL) and the total number of test takers who took each item at low score group (SL) were recorded. The difficulty index was calculated for the low score group (DL) as well. The high and low difficulty indices were then calculated to arrive at the discrimination index of each item. Out of 1,084 test items, 413 had an acceptable discrimination index above .30 (see Table 4).

Table 4
Item Analysis: Sample Discrimination Index

Item ID	Total no. of students	Facility value	Total no. of correct answers at high level (CH)	Total no. of students at high level (SH)	Difficulty index for high score group (DH = CH/SH)	Total no. of correct answers at low level (CL)	Total no. of students at low level (SL)	Difficulty index for low score group (DL = CL/SL)	Discrimination index (DH-DL)	Acceptable discrimination index: above .30 (√)
767/i4	*14*	*0.85*	*7*	*3*	*0.88*	*5*	*6*	*0.83*	0.04	
767/i2	*13*	*0.84*	*6*	*3*	*0.75*	*5*	*5*	*1.00*	-0.25	
640/i2	*34*	*0.79*	*13*	*13*	*1.00*	*14*	*21*	*0.67*	0.33	√
767/i3	*14*	*0.78*	*7*	*3*	*0.33*	*4*	*6*	*0.67*	0.21	
556/i4	*23*	*0.78*	*11*	*12*	*0.92*	*7*	*11*	*0.64*	0.28	
548/i1	22	0.68	10	11	*0.91*	5	11	*0.45*	0.45	√
556/i1	24	0.66	10	12	*0.83*	6	12	*0.50*	0.33	√

610/i1	47	0.65	18	23	*0.78*	13	24	*0.54*	0.24	
542/i1	22	0.63	7	12	*0.58*	7	10	*0.70*	-0.12	
649/i4	23	0.60	7	8	*0.33*	7	15	*0.47*	0.41	√
555/i1	29	0.58	12	13	*0.92*	5	16	*0.31*	0.61	√
547/i1	19	0.57	8	10	*0.30*	3	9	*0.33*	0.47	√
558/i2	21	0.57	8	14	*0.57*	4	7	*0.57*	0.00	
763/i1	30	0.56	11	15	*0.73*	6	15	*0.40*	0.33	√
558/i5	30	0.56	12	14	*0.36*	5	16	*0.31*	0.54	√
555/i5	38	0.55	13	17	*0.76*	8	21	*0.38*	0.38	√
			Total number of test items with acceptable discrimination index							413
			Total number of test items							1,084

The test items that fell within the .25 to .80 range of facility value and the items among them that had the highest discrimination index (>0.30) were further selected for inclusion in the revised test. This resulted in retaining only those items in the test that were well centered and discriminated well between the high and the low scoring students (see Table 5).

Table 5
Item Analysis: Sample Facility Value and Discrimination Index

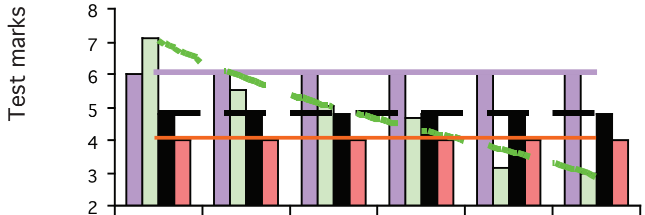
Item level	Item ID	Total number of students	Facility Value (FV)	Discrimination index	Items well centered and discriminating well: items with FV between 25% and 80% and with discrimination index above .30	Items not well centered and not discriminating well: items not with FV between 25% and 80% and not with discrimination index above .30
BENG-90	133/i1	39	0.84	0.26		√
BENG-90	118/i1	38	0.84	0.35	√	
BENG-90	720/i3	6	0.83	0.20		√
BENG-90	143/i4	6	0.83	0.50	√	
BENG-90	144/i1	6	0.83	0.20		√
BENG-90	144/i4	6	0.83	-1.00		√
BENG-90	648/i3	6	0.83	0.50	√	
BENG-90	290/i1	17	0.81	0.43	√	
BENG-90	719/i3	16	0.81	0.31	√	
BENG-90	269/i1	31	0.80	0.22		√
BENG-90	319/i5	25	0.80	0.11		√
BENG-90	145/i1	15	0.80	0.16		√
BENG-90	140/i3	5	0.80	-0.50		√
BENG-90	280/i1	27	0.77	0.17		√
BENG-90	312/i4	22	0.77	-0.18		√
BENG-90	272/i1	17	0.76	0.26		√
	Total of items with good FV and DI				390	
	Total of items with bad FV and DI					694
	Total number of test items					1,084

Three hundred and ninety items (35.98%) out of 1,084 were found to be well centered and discriminated well, and 694 items (64.02%) were found neither to be well centered nor good discriminators. In total, 390 items were found to have good facility value as well as good discrimination index. The rest of the test items were reviewed.

CUT-OFF POINTS

Because of the hybrid nature of the test, we also had to find a somewhat different system to arrive at cut-off points. The NEPTON test cut-off points were calculated for each level-based slide, and were the result of a long series of iterations on student scores from the field testing. First, we analyzed the NEPTON raw data, which indicated the average of all test takers taking all questions. We added the overall averages of all students taking all questions at each of the six levels and divided them by six to come up with the mass average (the total of the overall average). We made the resulting figure of 4.76 into an integer, 5, because we had to have whole numbers as cut-off points. It was decided that the upper and lower bounds would be one above and one below, that is, the upper bound would be 6 and the lower bound would be 4 (see Figure 9).

Figure 9
Bounds on Cut-off Points



	LEV1	LEV2	LEV3	LEV4	LEV5	LEV6
🟪 UPPER BOUND	6	6	6	6	6	6
🟩 OVERALL AVERAGE	7.12	5.52	5.05	4.67	3.18	3.00
⬛ MASS AVERAGE	4.76	4.76	4.76	4.76	4.76	4.76
🟥 LOWER BOUND	4	4	4	4	4	4

Once we found the upper and lower bounds, we started the iteration process to find acceptable cut-off points which would indicate the group average. Iteration 1 produced the lower bound cut-off points. These cut-off points were four at each of the six levels. Iterations were then carried out to try to refine the cut-off points. The result of Iteration 2 was 5, 4, 4, 5, 4, and 5. This result was again used in a series of iterations to see if we could come up with better cut-off points. In the end, it was found that the best cut-off points were the ones found in iteration 2, that is, 5, 4, 4, 5, 4, and 5 because they fell within the original boundaries, four for the lower bounds and six for the upper bounds.

As a result, the NEPTON placement was computed statistically and based on the principles of the nine-item/six-level slide design and the iteration process described above. Test takers moved from one level slide to another if they reached the designated cut-off point of each slide. In other words, if test takers reached the cut-off point of 5 for BENG-50, they moved on to the next level. If they reached the cut-off point of 4 for BENG-80, they moved on to the next level, and so on. This formula calculated the final placement and converted it to an Intercollege Placement English competence level.

FIRST NEPTON TEST ADMINISTRATION

The online version of NEPTON was administered in September 2004, at the beginning of the new academic year. The aims were to test and, if necessary, adjust the cut-off points derived from the iteration of the pen-and-paper field testing data and also to test the online delivery of the test. Since NEPTON is a test of grammatical and vocabulary knowledge as well as reading comprehension, there was a potential to misjudge student abilities compared to the written component, which is based on multiskill assessment. We therefore decided to trial NEPTON in its online format conjointly with the hand-written writing component. This accepted practice (Dunkel, 1999) was also necessary to implement in this project because it facilitated the introduction and easier acceptance of this new testing system by the English language instructors.

Five sessions were organized in eight 24-seat capacity IBM computer labs. A total of about 800 test takers were supervised by a trained invigilator per lab and serviced by lab assistants, the IT coordinator, the project coordinator and her assistant, and the head of the department.

Test takers were asked to complete pre- and posttest questionnaires: 263 test takers volunteered to complete the pretest questionnaire and 113 test takers chose to complete the posttest questionnaires.

FINAL NEPTON SCORE AND RESULT REPORTING

Test takers were placed in one of the six language levels by NEPTON only if they reached a minimum level at each level, according to the cut-off points previously calculated. The reported score indicated the English language level where test takers were placed. Both electronic and written results were considered, and the placement level of English of each test taker was determined.

Writing

All written tasks were based on tasks drawn from and reflecting meaningful written text-based communication that test takers would encounter during their college life as well as similar to ones found in the teaching materials of the respective levels. Test takers were given a selection of descriptive, narrative, informative, or persuasive written tasks of 120 words in length. The written task measured students' writing skills in an integrative way.

NEPTON'S RELIABILITY AND VALIDITY

Test Reliability

After designing the hybrid model and developing the test, we wanted to find out whether such a test could be an efficient, reliable, and valid testing tool for our large number of English test takers.

General factors influencing reliability

A fundamental concern in the development and the use of a language test is to minimize the effects of sources of unreliability (Bachman, 2003). We used Hughes' (1989) suggestions and

Dunkel's (1999) factor categorization to ensure and check NEPTON's reliability. NEPTON has a large item pool (635 questions, total of 1,084 items). Instructions are simple and clear. Data analysis indicated that most test takers (78.8%) felt quite comfortable with the instructions. In NEPTON, there is no choice of questions. The review process helped eliminate ambiguous items, reach agreement on acceptable responses, and edit questions and texts. Discrete objective scoring items are also used. Moreover, all items are randomly chosen and form a unique test for each test taker. In addition, the test specifications provide a detailed description of the system of cut-off points. The computer screen layout is clear and simple, and, as indicated by the data analysis, most test takers were satisfied with the clarity of both the computer screen (39.8% agree and 44.22% strongly agree) and the activity layout (36.3% agree and 35.4% strongly agree).

NEPTON's interitem consistency: split-half reliability index

After studying the various methods of measuring reliability (test-retest reliability, parallel-form reliability, split-half reliability index, KR 20 formula, and KR 21 formula), we decided to estimate NEPTON's reliability by measuring the interitem consistency. We simulated the parallel forms method by calculating the split-half reliability index. This involved dividing the test into two, using the odd-even method for splitting the items, treating these two halves as being parallel versions, and correlating these two halves. As Table 6 shows, the two halves of the NEPTON test correlated strongly, thus suggesting high reliability.

Table 6
Pearson Correlations (N = 866)

	Odd number	Even number
Odd number		.822*
Even number	.822*	

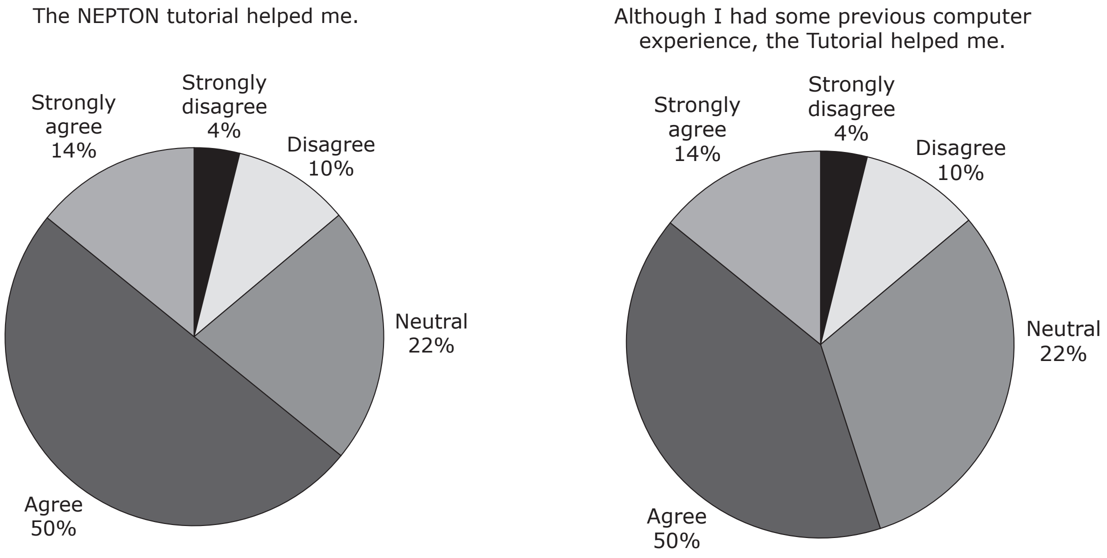
*p < .01 (2 tailed)

The item bank construction was based on needs analysis, the English language program review, and test specifications. As a result, we had a general expectation that the cut-off points would be around 4-5. Indeed they were. We also had the pen-and-paper cut-off points as a starting point. In addition, we studied the gross proportion of statistics from previous years which indicated the percentage of students placed at each level. There was no indication that there were differences in this year's intake. Moreover, we compared the electronic placement with the written component placement results to explore how they correlated. The total of 78% agreement (33% same placement and 45% one level difference) indicates a good correlation between the two components of the test, although a higher correlation of the same placement would have been more satisfactory. Twelve percent had two levels difference. In this case, test takers were placed in the middle level by a individual decision on those two students. The disparity of placement level—usually by one level—of 15% of the test takers was of concern and indicated the need for further investigation.

General and individual factors influencing reliability

Test takers need to be familiar with the format of a test before taking the test because familiarity makes test takers feel more comfortable. Test-taker factors such as transient factors (e.g., the physical and psychological health of the test taker) and stable factors (e.g., their experience with similar tests) may influence test takers' performance. For these reasons, test takers were given the opportunity to familiarize themselves with the test format and types of activities in the form of a booklet they received upon registration for the NEPTON test and also in the form of an electronic tutorial they completed just before taking the test. Test takers were asked how they felt before and during the test. Based on the data analysis, test takers' tension rate before the test averaged between 24.8% neutral, 28.8% agreeing, and 12.4% strongly agreeing. On the other hand, they expressed that they felt quite comfortable during the test, (27.4% agreeing and 29.2% strongly agreeing). Test takers were also asked whether the tutorial helped them (see Figure 10).

Figure 10
The NEPTON Tutorial and Computer Familiarity


The NEPTON tutorial helped me.

Although I had some previous computer experience, the Tutorial helped me.

Sixty-four percent (14% strongly agreed and 50% agreed) stated that the tutorial was helpful to them. Even the majority of those who had some previous computer experience strongly agreed (14%) or agreed (41%). Although a considerable number of test takers were neutral, disagreed or strongly disagreed about this, only three out of more than 800 opted to take the test in its pen-and-paper format. This suggests that the majority of test takers preferred to take the placement test online.

Situational factors influencing reliability

The conditions of the test administration were uniform and nondistractive: test takers were sent to specific computer laboratories, and invigilators helped students through the tutorial. English language test experts and lab assistants were available at all times. There were no major distractions for most test takers, apart from one occasion in which a system failure, due to excessive demand of the system, caused a crash—a problem that was dealt with quickly.

Test Validity

We then checked the extent to which the objective component of NEPTON was sufficiently valid for its purpose and also the reactions and feelings of test takers and faculty towards it. An adequate sample (more than 800 test takers) was used in the test trial for the test validation. Two hundred sixty-three answered the pretest questionnaire and 113 answered the posttest questionnaire. They are representative of the population for which the test is intended in age, experience, and background. The language levels of the test provide an adequate basis for validating the instrument. The large size of the item pool (1,084 items) also supported higher test validity.

Internal face validity

First, we established the extent to which the test takers were familiar with computers and examined their attitudes and feelings towards the computer delivery of NEPTON. We wanted to find out whether the eventual English language placement would be significantly affected by the test takers' familiarity with computers.

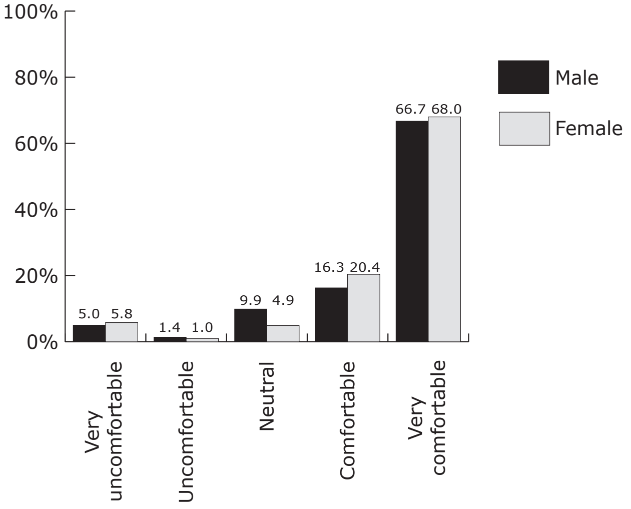
Test taker choice of the test mode of delivery

Test takers had a choice of taking the test electronically or in pen-and-paper format. Out of more than 800 students, five students said they wanted to take the paper version of NEPTON, but, after doing the NEPTON tutorial, only three ultimately chose this version of the test.

Test taker computer familiarity

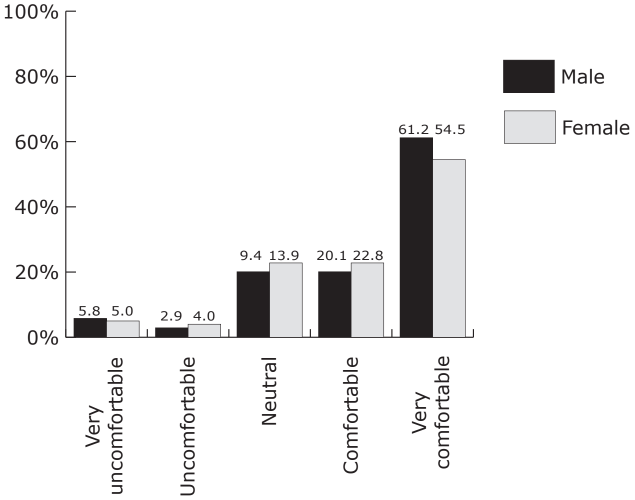
Before taking NEPTON, data were collected from each test taker on computer familiarity. The data were analyzed to establish test takers' prior computer familiarity and attitudes toward taking a test electronically in comparison with the pen-and-paper test. This involved asking test takers whether they felt comfortable with basic computer techniques such as mouse clicking (see Figure 11) and scrolling (see Figure 12).

Figure 11
Test Taker Familiarity with Basic Computer Skills: Mouse Clicking



Sixty-six point seven percent of the male and 68% of the female test takers felt very comfortable with mouse clicking; only 5% of the males and 5.8% of the females felt very uncomfortable, and 1.4% of the males and 1% of the females felt uncomfortable with mouse clicking.

Figure 12
Test Taker Familiarity with Basic Computer Skills: Scrolling



Sixty-one point two percent of the males and 54.5% of the females felt very comfortable with scrolling. On the other hand, 5.8% of the males and 5% of the females felt very uncomfortable, and 2.9% of the males and 4% of the females felt uncomfortable with scrolling.

From these data, it was evident that many test takers felt adequately comfortable using basic computer techniques (mouse clicking and scrolling) required for NEPTON, but there was a considerable number who did not seem to be. On the whole, the test takers did not seem to feel disadvantaged by the use of technology. On the contrary, as we have seen, the majority preferred to take the electronic version. Another indication was many test takers' preference for the electronic test when they were given the hand-written task. However, we need to investigate more carefully the case of those test takers who feel uncomfortable with the basic computer techniques they need in order to take the test. At a later stage, it is planned to develop, in addition to the tutorial, a nonsecure component of the test in which prospective test takers can practice taking the test online at their leisure weeks or even months before actually taking the test.

Test takers' impressions

Based on the pre-NEPTON test taker questionnaire data analysis, most test takers found the different types of questions covering the different skills to be manageable. The majority also found the instructions clear (43% agree and 46% strongly agree). The test takers also found the topics of the test interesting (38% agree and 23% strongly agree), while 22.1% were neutral, 8.0% disagreed, and 2.7% strongly disagreed. The test takers also found that there was enough variety of activities (46% agree and 15.9% strongly agree), but 23.9% were neutral, 9.7% disagreed, and 1.8% strongly disagreed.

As a whole, the results above indicate a general acceptance of and a feeling of comfort with the NEPTON test. The test takers' attitudes and reactions indicate a general acceptability of items and test components.

Internal content validity

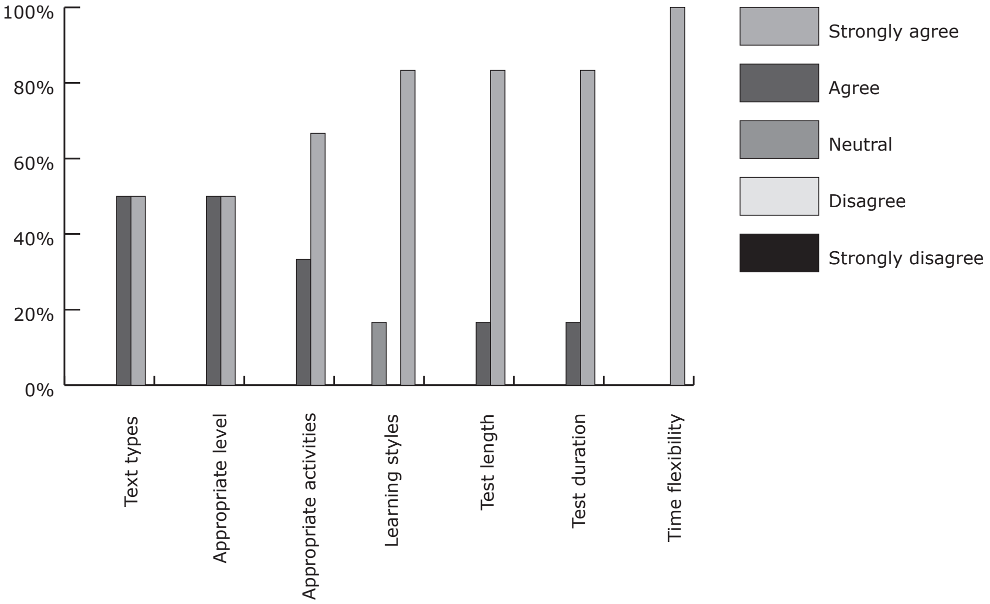
Six experienced English lecturers examined NEPTON's content validity by studying two sample pen-and-paper tests and a sample electronic test and comparing all three tests to the Test Specifications. They rated NEPTON using a questionnaire. The results are discussed below.

The data analysis shows that 33.33% of the experts strongly agreed and 66.66% agreed that NEPTON met the requirements of an English placement test for the needs of University of Nicosia (Intercollege) students. Fifty percent strongly agreed and the other 50% agreed that NEPTON accurately represented the content of the University of Nicosia (Intercollege) English language program at the various levels.

In addition, 50% strongly agreed and 50% agreed that the grammatical points represented the content of the respective levels of instruction. Sixty-six point sixty-six percent strongly agreed and 33.33% agreed that the lexical points represented the content of the respective levels. Sixty-six point sixty-six percent strongly agreed and 33.33% agreed that the reading comprehension sections represented the content of the respective levels. Sixteen point sixty-six percent strongly agreed and 33.33% agreed that the topics were appropriate for the respective levels, while another 16.66% were neutral. The majority of experts (66.66%) agreed that the topics were interesting, whereas 16.66%, respectively, strongly agreed or were neutral. All experts strongly agreed that the instructions were clear.

The experts' opinions varied in the following areas: 33.33% strongly agreed that the test reflected the test takers' ethnic background, 16.66% agreed and 33.33% disagreed. Sixty-six point sixty-six percent strongly agreed that the test reflected settings test takers would encounter in their studies, whereas 16.66% agreed and another 16.66% were neutral. Eighty-three point thirty-three percent strongly agreed that the test reflected settings and contexts in English-speaking environments. Another 16.66% felt neutral. Finally, 83.33% strongly agreed that the test reflected the setting in which test takers would be studying, whereas 16.66% felt neutral about this issue (see Figure 13).

Figure 13
Experts NEPTON Content Validity Evaluation



Fifty percent of the experts strongly agreed and another 50% agreed that there was sufficient variety in text types and that the NEPTON test items generally targeted the appropriate level. Sixty-six point sixty-six percent strongly agreed and 33.33% agreed that the kinds of activities were appropriate. Eighty-three point thirty-three percent strongly agreed that the activity types covered a variety of learning styles; on the other hand, 16.66% felt neutral.

Eighty-three point thirty-three percent of the experts strongly agreed and 16.66% agreed that the time given to the test takers to take the test was appropriate, and all of them strongly agreed that there was enough flexibility in time to cater to individual differences in test takers' time needs.

The experts were then asked to evaluate specific aspects relevant to the electronic component of NEPTON only. According to the data analysis, 83.33% strongly agreed and 16.66% agreed that the tutorial helps test takers familiarize themselves with the computer skills they need to take the NEPTON test, even those with some previous experience with computers. All experts strongly agreed that the tutorial helps the test takers familiarize themselves with the NEPTON test format, layout, and test items.

All experts strongly agreed that the computer screen and the activity layout were clear and that test takers are given opportunities to control the test (e.g., review items and go backwards and forwards).

Figure 14
NEPTON's Efficiency

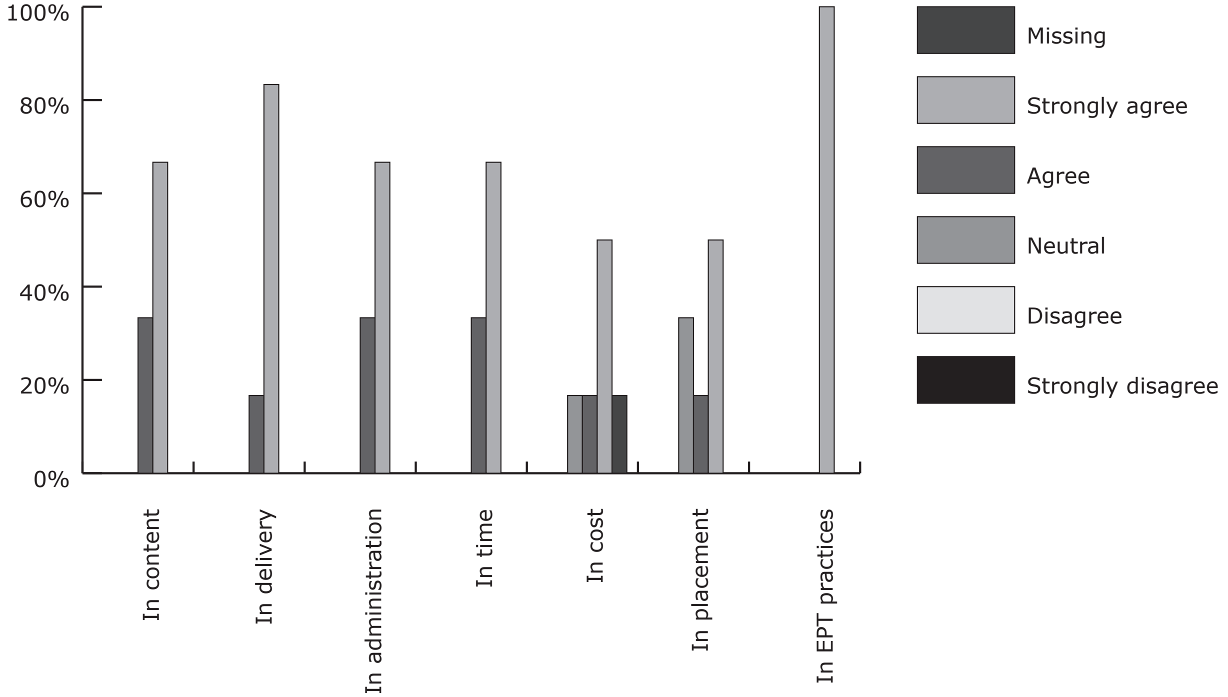


Figure 14 shows that 66.67% strongly agree that NEPTON is more efficient in content, administration, and time. Eighty-three point three percent strongly agree and 16.7% agree that NEPTON is more efficient in delivery mode. Fifty percent strongly agree that it is more efficient in cost and placement. Sixteen point seven percent agree on both, while 16.7% are neutral about the test's efficiency in cost, 6.7% chose not to answer about the cost and 33.3% were neutral about whether NEPTON was more efficient in placement. All experts strongly agreed that NEPTON has brought improvement in English placement test practices.

Figure 15
NEPTON Better than Previous English Placement Test

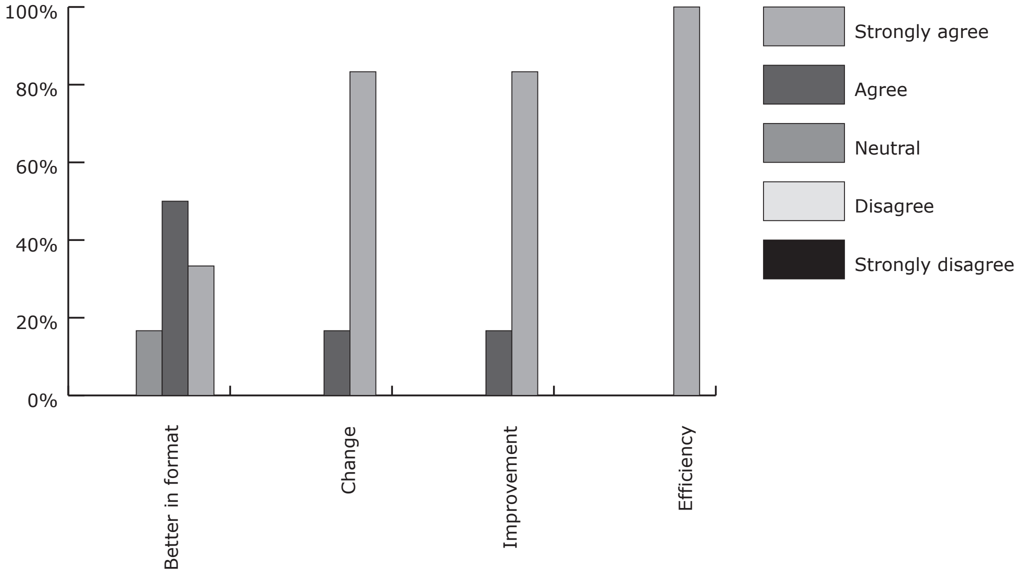


Figure 15 shows that 16.7% felt neutral, 50% agreed, and 33.33% strongly agreed that NEPTON was better in format than the previous English placement test. Finally, 83.3% of the experts strongly agreed and 16.7% agreed that NEPTON had brought change and improvement, and all of them agreed that it brought efficiency in University of Nicosia (Intercollege) practices.

Conclusions from the experts' test evaluation

Many of the aims of the research project seemed to satisfy the experts, aspects such as variety in text types, interesting topics, clarity, efficiency, user friendliness, security, format, review capabilities, and time. The experts' evaluation indicates that in general NEPTON offers improvement, change, and efficiency in the English placement practices at the University of Nicosia (Intercollege). It can also be inferred that NEPTON reflects the test specifications and current theories and practices in L2 testing. Their suggestions for improvement positively contribute to its further improvement.

External construct validity

The construct validity of a test is "the extent to which the test may be said to measure a theoretical construct or trait" (Anastasi, 1982, p. 144). In the case of NEPTON, the agreement of experts in the field regarding the content validity of the test also lent credence to the inferred degree of construct validity. English instructors at the University of Nicosia (Intercollege) were also asked to give their feedback about the NEPTON. They were happy to use it and replace the old placement test, but they wanted to keep the written component, even though this is not offered electronically for the time being.

CONCLUSIONS

In Cyprus, there are a number of public and private tertiary institutions. Some of these institutions feature English as the language of instruction. For this reason, they have English placement tests to place their students in their English language courses. All of them, however, use pen-and-paper English placement tests. The NEPTON test at the University of Nicosia (Intercollege) is the first English placement test used in Cyprus at tertiary level in an electronic form and is based on contemporary theories and practices in L2 testing. This article documents the processes followed to arrive at a choice of the university's testing model and the design of a hybrid online testing model felt to be the most suitable for the English placement testing practices of the university. The discussion of the test implementation and evaluation also provides evidence of the extent to which the hybrid NEPTON proved to be efficient and reliable as a testing tool for large groups of English learners in a particular setting. The NEPTON test model should not be seen as a static and permanent means of placement test practices at the University of Nicosia (Intercollege). One of the reasons for developing an in-house model was to have the opportunity to continuously monitor, develop, and change the model according to changes in the area of technology-based language testing and test takers' characteristics. Also, this model should not be seen as a model that could simply be transferred to other English placement testing programs because it was developed for a specific context and specific needs. Its case (alternative design and innovative features), however, may prove useful to other English placement testing programs.

REFERENCES

The TOEFL Test Details: Learners and Test Takers Internet-based Testing (iBT). (n. d.). Available from http://www.ets.org

A common European framework of reference for languages. Learning, teaching, assessment. (n. d.). Retrieved October 8, 2007, from http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

Alderson, J. C. (1987). Innovation in language testing: Can the micro-computer help? [Special Report No. 1]. *Language Testing Update*. University of Lancaster: Lancaster, UK.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. New York: Cambridge University Press.

Almond, R. G., Steinberg, L. S., & Mislevy, R. H. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning and Assessment, 1* (5), 3-63. Retrieved October 8, 2007, from http://escholarship.bc.edu/jtla/vol1/5

Anastasi, A. (1982). *Psychological testing* (5th ed.). New York: Macmillan.

Bachman, L. F. (2003). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Blackboard. (2004). *Blackboard learning system*. Washington, DC: Blackboard, Inc. Available from http://www.blackboard.com

Brown, J. D. (1997). Computers in language testing: Present research and some future directions*. Language Learning & Technology, 1* (1), 44-59. Retrieved October 8, 2007, from http://llt.msu.edu/vol1num1/brown/default.html

Brown J. D. (June, 2003). Norm-reference item analysis (item facility and item discrimination). *Shiken: JALT Testing & Evaluation SIG Newsletter, 17* (2), 16-19. Retrieved October 17, 2007, from http://jalt.org/test/bro_17.htm

Cohen, A. (1984). Fourth ACROLT meeting on language testing. *TESOL Newsletter, 18*, 23*.*

Chapelle, C. A. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing and research*. Cambridge: Cambridge University Press.

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through technology*. Cambridge: Cambridge University Press.

DIALANG. (2001). Available from http://dialang.org./intro.htm

Dunkel, P. (1999). Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning & Technology, 2* (2), 77-93. Retrieved October 8, 2007, from http://llt.msu.edu/vol2num2/article4

Fulcher, G. (2000). Computers in language testing. In P. Brett & G. Motteram (Eds.), *A special interest in computers* (pp. 93-107)*.* Manchester: IATEFL Publications. Retrieved October 8, 2007, from http://www.le.ac.uk/education/testing/ltrfile/Computers.html

Fulcher, G. (2003). Interface design in computer-based language testing. *Language Testing, 20* (4), 384-408.

Godwin-Jones, R. (2001). Language testing tools and technologies. *Language Learning & Technology*, 5 (2), 8-12. Retrieved October 8, 2007, from http://llt.msu.edu/vol5num2/emerging/default.html

Grist, S. (1989). Computerized adaptive tests (ERIC Digest No. 107). ERIC Clearinghouse on Tests Measurement and Evaluation. Washington DC: American Institutes for Research. Retrieved October 8, 2007, from http://www.ericdigests.org/pre-9213/tests.htm

Half-Baked Software. (n. d.). *Hot potatoes* (Version 6.0.3). Victoria, British Columbia, Canada: Half-baked Software, Inc. Available from http://www.halfbakedsoftware.com

Heaton, J. B. (1988). *Writing English language tests.* New York: Longman.

Hughes, A. (1989). *Testing for language teachers.* Cambridge, UK: Cambridge University Press.

Item Analysis. (n. d.). East Lansing, MI: Scoring Office, Michigan State University. Retrieved October 8, 2007, from http://www.msu.edu/dept/soweb/itanhand.html

Kitao, S., & Kitao, K. (1996). Testing communicative competence. *The Internet TESL Journal, 2* (5). Retrieved October 8, 2007, from http://iteslj.org/Articles/Kitao-Testing.html

Larson, J. W., & Madsen, H. (1985). Computerized adaptive language testing: Moving beyond computer-assisted testing. *CALICO Journal, 2* (3), 32-37. Retrieved October 8, 2007, from https://calico.org

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Mack, D., & Seven, D. (2002). *Programming data driven web applications with ASP.NET*. Indianapolis, IN: SAMS Publishing.

Noijons, J. (1994). Testing computer assisted language tests: Towards a checklist for CALT. *CALICO Journal, 12* (1), 37-58. Retrieved October 8, 2007, from https://calico.org

Papadima-Sophocleous, S. (2005). *Development, implementation, and evaluation of an online English placement test at college level: A case study.* Doctorate of Professional Studies project, Middlesex University.

Quia Corporation. (1998-2006). *Quia*. Available from http://www.quia.com

Quick Placement Test [computer software]*.* (2001). Cambridge, Oxford, UK: University of Cambridge, Oxford University Press.

Question Mark Perception [computer software]*.* (n. d.). London, UK: Question Mark Computing Ltd.

Roever, C. (2001). Web-based language testing. *Language Learning & Technology, 5* (2), 84-94. Retrieved October 8, 2007, from http://llt.msu.edu/vol5num2/roever/default.html

Sceppa, D. (2002). *Microsoft ADO.NET.* Redmond, WA: Microsoft Press.

Stevenson, J., & Gross, S. (1991). Use of a computerized adaptive testing model for ESOL/bilingual entry/exit decision making. In P. Dunkel (Ed.), *Computer-assisted language learning and testing* (pp. 223-235). New York: Newbury House.

Statistics Guide, Interpreting the Reports (n. d.). Retrieved October 8, 2007, from http://helpdesk.kent.edu/howto/testscoring/stat

TOEFL Sampler [computer software]. (1998). Princeton, NJ: Educational Testing Service.

TOEFL Testing Program [Online]. Available from http://www.ets.org

Vispoel, W. P., Hendrickson, A. B., & Bleiler, T. (2000). Limiting answer review and change on computer adaptive vocabulary tests: Psychometric and attitudinal results. *Journal of Educational Measurement, 37* (1), 21-38.

Wainer, H., & Eignor, D. (2000). Caveats, pitfalls, and unexpected consequences of implementing large-scale computerized testing. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computer adaptive testing: A primer* (2nd ed.) (pp. 271-299). Hillsdale, NJ: Laurence Erlbaum Associates.

Walther, S. (2003). *ASP.NET unleashed* (2nd ed.). Indianapolis, IN: SAMS Publishing.

WebCT (Version 3.0) [online]. (2004). Washington, DC: Blackboard, Inc. Available from http://www.blackboard.com

Weir, C. J. (1990). *Communicative language testing*. Wiltshire, UK: Prentice Hall International.

ACKNOWLEDGMENT

Special thanks to Dr. Dmitry Apraksin for the NEPTON test system architecture in Figure 6.

AUTHOR'S BIODATA

Dr. Salomi Papadima-Sophocleous, Assistant Professor at the University of Nicosia (Intercollege), Cyprus, holds *DProf* in Applied Linguistics: Online English Language Testing, Middlesex; *LittM*: French Literature, UNE; *MEd*: L2 Curriculum Evaluation/Implementation; *PostGradDip* CALL, Melbourne; *DipEd*: L2 methodology: French/Greek, La Trobe; *GradCert*: TESOL; and *BA*, Athens: French/Greek. She was awarded the 2005 Ken Goulding Prize for Professional Excellence, Middlesex University. She teaches TEFL (including Testing), CALL, Academic English, Greek, and French. She has extensive experience in teaching at secondary and tertiary levels and teacher training, on campus and online (La Trobe and RMIT universities, Australia, University of Nicosia (Intercollege), Cyprus), assessment (Australian Victorian Assessment Programme), curriculum/test development (printed and electronic), programme evaluation, CALL, and in International and European programs (E.D.I.A.MM.E).

AUTHOR'S ADDRESS

Dr. Salomi Papadima-Sophocleous
Department of Languages
School of Humanities, Social Sciences, Law and Languages
University of Nicosia (Intercollege)
46 Makedonitissas Avenue
P.O. Box 24005
1700 Nicosia
CYPRUS
Phone: + 357 22 841 636
Fax: + 357 22355 166
Email: papadima.s@unic.ac.cy