Difficulties and Constraints Involved in Developing a New English Placement Test Online

Dr. Salomi Papadima-Sophocleous

Intercollege, Cyprus

papadima.s@intercollege.ac.cy

Abstract

The purpose of this work based research project was to explore ways to bring change, improvement, and efficiency into specific college-level English placement practices based on current theories and practices on the one hand, and on the specific contextualised student needs and the local human and other resources on the other. The comparison of theoretical suggestions in online English placement testing with the English placement testing (EPT) practices at a particular college in Cyprus, helped in shaping its New English Placement Test Online (NEPTON) specifications, and deciding on the test type and design, method of test development and implementation. The combination of current theories and practices with the potentiality of putting them into practice in the best possible way within college resources, resulted in choosing to develop a computer based test (CBT), with some computer adaptive test (CAT) features such as large test item pool, test item randomisation, cut-off point algorithm, and an alternative way to test item analysis as a result. Moreover, the test design had to also include the design of an in-house electronic environment for hosting and delivering the placement test and facilitating continuous test monitoring and improvement. The data analysis carried out to establish test reliability and validity shed light on the strengths and weaknesses of such a research project and complemented the description of the processes of EPT creation based on theory but mainly the description of the difficulties and the constraints deriving from the particular context and its human and physical resources.

Keywords: English Placement Test Online; resources constraints

Introduction

According to current theories (Hughes, 1989; Weir, 1990, 1995; Alderson, Clapham, & Wall, 1995; Heaton, 1998; McNamara, 2000; Council of Europe, 2001; Chalhoub-Deville, 2001; Fulcher, 2000; Godwin-Jones, 2001; Kitao, & Kitao, 2002; Roever, 2001; Chapelle, & Douglas, 2006) and practices (French Cat, 1995; CB IELTS, 2005; TOEFL CBT, 1998; QPT, 2001; TOEFL iBT, 2005) during the last decade, language testing has been shifting from pen and paper to electronic testing with the aim of addressing efficiency and effectiveness (Grist, 1989, Dunkel, 1999) in placement, delivery, time, costs and human resources. Throughout the world we have seen examples of such tests developed (McNamara, 2000, Chap. 8, plus Text 20 pp. 116-119) either by large academic or commercial institutions for large numbers of students (TOEFL Testing Program, 2001, Quick Placement Test, Oxford and Cambridge Universities, 2001), or by a handful of testers who aim to cater for the testing needs of their particular setting (the Monash / Melbourne French CAT, 1995). The first type has several advantages: large amounts of money or grants to support them, usually a team of experts from different disciplines working to develop such tests, and sufficient timeframe to trial, implement and evaluate them. Small groups, however, face the challenge of combining current theories with application realities which include restricted human and other resources.

For many years, at the college where I work, a private English-speaking university in Cyprus, the English Placement Test was delivered in a pen-and-paper form. In 2003, we decided to improve our practices and develop a New English Placement Test Online (NEPTON) (NEPTON, 2005; Papadima-Sophocleous, 2005). The requirements from the administration were for this task to be achieved based not only on current theories and practices in Computer Assisted Language Testing (CALT), and the specific contextualised student needs, but also on the local human and other resources for the test development, test administration and the reporting of results (A.L.T.E. 2002, p. 8). A constraint is a restriction

on the degree of freedom one has in providing a solution, in this case develop a test. Constraints are effectively global requirements, such as limited development resources or a decision by senior management that restricts the way one develops a system. Constraints can be economic, political, technical, or environmental and pertain to one's project resources, schedule, target environment, or to the system itself. This paper describes how the NEPTON test construction was achieved within such constraints.

Constraints

The realities of the implementation of such a project were very tight. Human resources were limited: The project was assigned to a language staff member, with many years of experience and expertise in language teaching and learning, curriculum development, testing, and Computer Assisted Language Learning, mainly in Australia and also in Cyprus. This staff member was given three hour weekly time release for three semesters. Two other staff members formed the project team: an experienced computer science programmer, who was given a small money allowance, and a second language staff member, who worked on a voluntary basis. Any involvement of any other people was strictly above his or her other academic duties. About 2000 students participated, (1200 in the field testing and 800 in the test trial), and 70 staff members acted as moderators, invigilators, statisticians, administrators, or lab assistants.

Urgency, Time, Monetary, Hardware and Software Constraints, and Expertise

The urgency of the project was another constraint. It had to be developed within a year. This deadline was extremely tight and made the process and decision making quite difficult. The time available for the team to work on it as a team and as individuals was a serious constraint. As mentioned before, only the language staff member who was assigned the

project was given some time release (only three hours per week for three semesters, including summer session). She was simultaneously involved in teaching, research publications, and her doctorate studies. The second language staff member worked on the project on a voluntary basis, in addition to his teaching, research and Language Laboratory coordinating duties, and his doctorate studies. Many times, meetings had to be postponed and work had to be cancelled due to other urgent matters that arose and which needed more immediate attention by one or the other member of the group. Some monetary allowance was given only to the computer programmer to design, develop and implement the electronic delivery of the test. He was also simultaneously teaching, leading the Computer Centre, and coordinating all computer activities for all three campuses. Software and hardware needed for the project had to be supplied from within the institution existing resources. These constraints put restrictions on the team and pressure to find alternative ways to implement the project.

Gaining extra expertise fast was another challenge faced by the team. For example, the computer programmer with years of experience in programming had never worked on developing a language test. The project leader had extensive experience as an L2 examiner and a chief examiner and in computer assisted language learning but had never combined these two to create an online language placement test. Statistics to carry out item analysis for reliability and validity, and to establish cut off points were also needed for such a test, something she had little knowledge of, since most of her previous research was mainly of a qualitative nature. The hybrid nature of the test also added more challenge for new knowledge. The second language staff member had extensive experience in language testing and the use of New Technologies and some knowledge of statistics. However he had to expand his knowledge in this area for this project.

The above brief description of the project's human and other resources clearly indicates the constraints the project team had to work with to develop the NEPTON test within the practicalities of our institution.

Project phases

The project development phases included:

Literature Review in the area of Online Test Development

Needs analysis

Test specification development (determination of type of test and development of test content and electronic feature specifications)

Item bank writing

Test item moderation

Item selection algorithm creation

Testing shell design and implementation

Test item uploading

Field testing

Data and item analysis and test item improvements

Data analysis to derive cut off points for the electronic test algorithm

Test trial

Data analysis for test reliability and validity

Each team member took on tasks according to his or her expertise, project needs and availability.

(a) The project leader carried out the literature review and the needs analysis (West, 1994; Brown, 1995, p.35; Witkin, & Altschuld, 1995), designed the new test specifications, both for the content and the test delivery software, wrote the test items, coordinated the test moderation, made the resulting changes and uploaded test items, organised and

managed the test trial and implementation, ran the item analysis, developed the item selection with the second language faculty, and the cut off point algorithms with the second language staff member and a volunteer statistician; developed the test electronic tutorial, trial test and test printed sample; wrote the invigilators' instruction booklet; ran the data analysis for the test reliability and validity, with some help from a second volunteer statistician.

(b) The computer programmer designed, developed and monitored the implementation of the electronic test delivery environment, based on the English language test design and specifications.

(c) The second language staff member and English language instructor helped with some of the test item writing, contributed to the item selection algorithm development, and worked heavily on the test-takers' data analysis and iteration which helped derive the cut off point algorithm. He also contributed to the moderation, the test field-testing organization, and the test trial.

Despite all the constraints mentioned above, the team managed to develop the NEPTON test, within the existing local human and other resources and based on current theories and practices in Computer Assisted Language Testing (CALT), and the specific college student needs. The next section gives a brief description of the project phases.

Needs Analysis

The Existing EPT

The examination of relevant documents and the existing EPT practices (Papadima-Sophocleous et al., 2005) indicated that these were not based on any clearly defined criteria based on specific theoretical grounds. It was mainly developed ad hoc and based on years of experience of instructors and testers of the institution. The examination indicated that

improvements were needed in the area of test planning and design, pass mark setting, test delivery, efficiency, based on current theories (Hughes, 1989; Weir, 1990, 1995; Alderson, Clapham, & Wall, 1995; Heaton, 1998; McNamara, 2000; Council of Europe, 2001; Chalhoub-Deville, 2001; Fulcher, 2000; Godwin-Jones, 2001; Kitao, & Kitao, 2002; Roever, 2001; Chapelle, & Douglas, 2006) and practices (French Cat, 1995; CB IELTS, 2005; TOEFL CBT, 1998; QPT, 2001; TOEFL iBT, 2005) in L2 testing, and the particular needs of the target testers.

English Language Test Choice

The project leader then reviewed the current theories and practices in L2 testing (Hughes, 1989; Weir, 1990; 1995, Alderson, Clapham, & Wall, 1995, Heaton, 1998; McNamara, 2000; Council of Europe, 2001; Chalhoub-Deville, 2001; Fulcher, 2000; Godwin-Jones, 2001; Kitao, & Kitao, 2002; Roever, 2001; Chapelle, & Daglas, 2006). Since it is often not possible to incorporate communicative elements in tests, the project leader decided to adopt a form of the 'loosely' used Communicative Paradigm (Kitao, & Kitao, 1996), as the approach for the design of the NEPTON test, which includes communicative elements in a test. This means communicative elements are incorporated in the test when communicative tasks cannot be included. Chosen text types are authentic (as they are used in real communication) or authentic like, similar to the ones students come across in their academic, personal and social settings in Cyprus and overseas. They cover topics, and incorporate vocabulary, structure, settings, contexts, and sociolinguistic elements which derive from the most current L2 theories, learning materials, and level descriptors, which reflect the various levels students would be placed in and the contextualized situations students would be likely to find themselves in. The test assesses writing using two different approaches: it tests vocabulary and grammar with the use of objective, electronic testing activities, in the form of

sentence-based, multiple-choice items, and in a more contextualized mode of texts with four or five multiple-choice questions of dropdown menu selections. Reading comprehension is tested in two forms: in the contextualised and situational form of signs (accompanied by visuals), and in texts-based items with multiple-choice questions. Each item is selected according to the test purpose, topic presented, skills tested, activity type and format, the sociocultural context, and the test-taker background and study setting. The test also assesses writing through a more direct, communicative and extended global integrative writing task, in a non-electronic mode, that is, hand written.

English Language Computer Based Practices and Computer-Based Test Choice

The language faculty then reviewed the various computer based test practices (Assessment electronic authoring tools of comprehensive learning online environments such as WebCT 2004, and commercial or free electronic assessment devices such as *Question Mark* n.d., *Hot Potatoes* 2004, and *Quia* 1998-2006). The examination of such tools revealed that, although they offer a variety of online testing techniques, they are generic, restrictive in their functions because they only provide templates of specific test activity types and, more importantly, they depend on central, external control and require a user fee. Some universities and colleges have developed their own electronic English Placement tests, (*Quick Placement Test* 2001). We decided to develop our own tool, based on the specific needs of our placement testing programme, compatible with our English language curriculum, controlled and monitored by us. Moreover, its development had to be within our existing human and other resources, and with minimal costs, as described earlier (see paragraph on constraints).

The NEPTON Test Model

The project leader then studied and compared computer based and computer adaptive tests (Henning, 1987; Weir, 1990, 1995; Brown, 1997; Bachman, 2003), and, in consultation with the rest of the team, designed a hybrid test. The NEPTON test Hybrid model incorporates some advantages from both the Computer Based Test (CBT) and Computer Adaptive Test (CAT) tests (Brown, 1997; Dunkel, 1999; Roever, 2001; Stevenson, & Gross, 1991; Chapelle, & Douglas, 2006) and avoids some of their limitations: Like the CAT and unlike the CBT where all items are the same for all candidates, each item is randomly selected according to set criteria during the test administration, however not from an item pool like in the CAT, but from sub pools based on six language performance levels, different language skills, and activity types. Unlike a CAT, it assesses candidate in all levels as a CBT to establish Test-taker's knowledge at all language levels. Like a CBT and unlike a CAT, each candidate is aware of the whole number of items he or she has to answer and can allocate his or her efforts accordingly. The Hybrid NEPTON model is long enough to provide adequate information to place the candidate more accurately, compared to the longer CBT and the shorter CAT. Although items are administered one at a time like the CAT, each candidate can browse through the items, skip some to be answered later in the test, and review and change answers, like the CBT. Like the CAT and unlike the CBT, each test is different and unique for each candidate, but based on a systematic item selection algorithm. Unlike the CAT, where the cut off points are based on Item Response Theory (IRT), resulting from a cut off point algorithm which adapts from level to level according to candidate's responses, and unlike the CBT where placement are the results of accumulated marks, the Hybrid model cut off points are a result of an iteration process of data analysis of student's results at all levels. The Hybrid model can be both fixed as the CBT and available at any moment suitable to the student as the CAT. Like the CAT and unlike the CBT, it can provide immediate results.

Moreover, and very importantly, the Hybrid NEPTON model fitted more within our time framework, our expertise, and human, financial and other resources.

Test Item Writing, Bank and Moderation

Item writing was based on the test specifications developed as a result of the needs analysis. A pool of about 750 questions, totalling more than 1500 items was originally developed for the six English course levels and served as an item bank. The items were widely chosen from the whole area of content (for all six levels) and were presented in various forms. About 42% of English programme-teaching staff volunteered to moderate them. The project leader coordinated this process, communicated electronically or in person on a continuous basis with all moderators, and made sure all was done within the time limits set. She then processed all moderators' input.

Development of Electronic Test Design and Testing Software

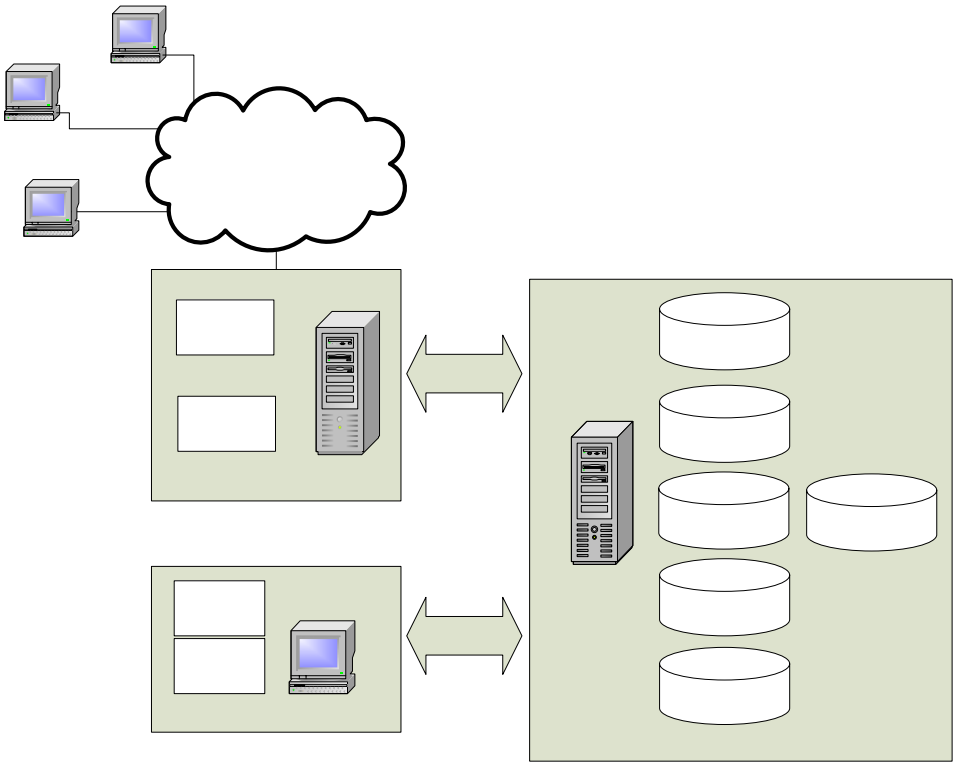
Meanwhile, based on the test specification document (Alderson et al., 1995; Kitao, & Kitao, 2002), including the hybrid test features, the computer programmer designed and developed an in-house English language online testing environment, using software and hardware within the existing resources of the institution. The system was developed on Microsoft.NET platform (Mack, 2002; Walther, 2003; Sceppa, 2002). Two Windows 2003 servers provide database and WEB server functionality. Microsoft SQL server 2000 (Nielsen 2003) is used as a system database server and native Internet Information Services (IIS) 6 (Tulloch, 2003) as a web server. The system architecture includes the following components and functions:

The Test Database Server hosts the Question Bank, the Test Profile, and the generated tests, the Students' Records, the Test History and the Assessment Rules.

The WEB server hosts the WEB Based Administrator and the Test Presenter.

The Desktop Administration Tool facilitates the test and test-taker record management.

Figure 1: The NEPTON Test System Architecture



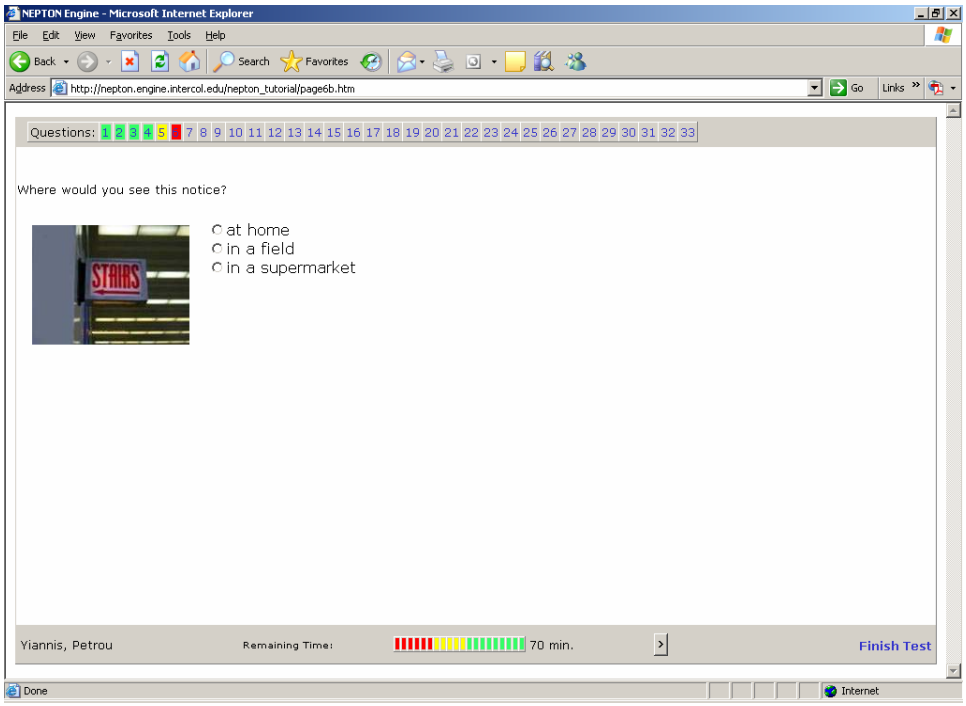
Both the test-taker and the administrator interface are user friendly, simple and kept to the minimum, even for those with minimal or no computer skills (mouse clicking and scrolling).

Test-taker Interface

The number of questions is clearly indicated at the top of the screen. The different button colours indicate the status of each item: green indicates questions already answered; yellow indicates that not all items in a text-based dropdown menu selection or multiple choice questions are completed, and that the test-taker needs to return and complete them; red indicates the item the test-taker is at; grey indicates items still remaining to be answered.

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.

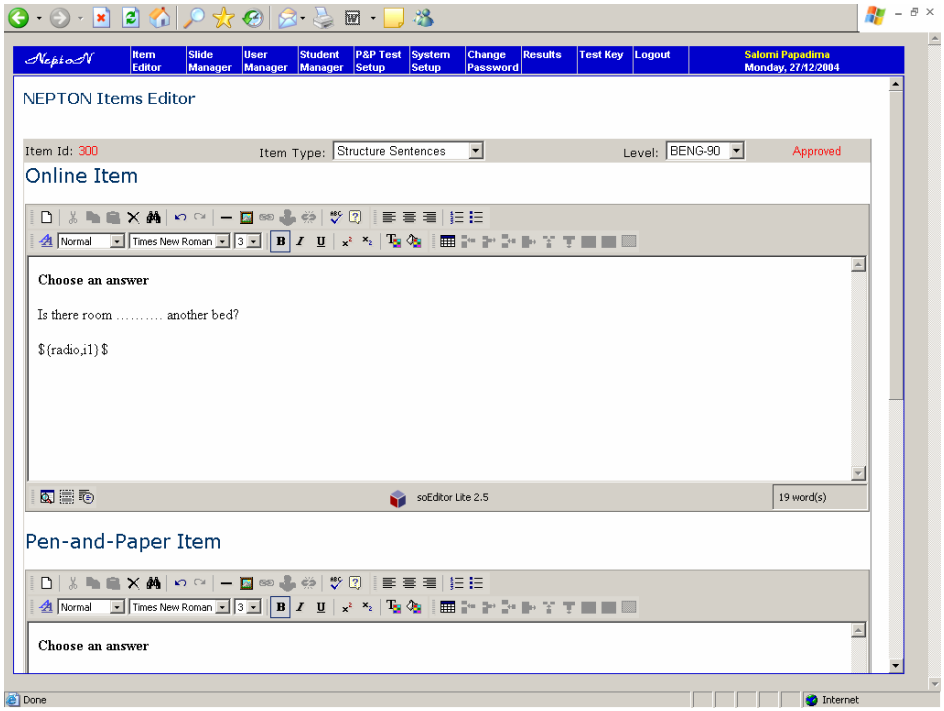
Figure 2: NEPTON Test Features

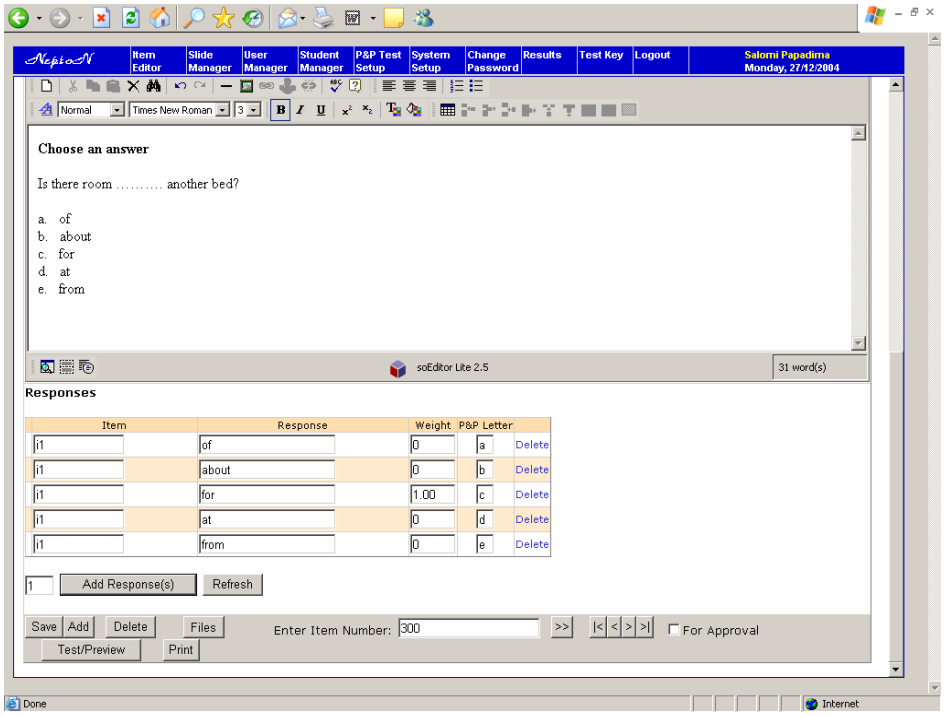


Administrator Interface

In the test-taker's interface, 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.

Figure 3: NEPTON's Item Editor





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 rest of the functions of the upper toolbar of the administrator's interface are equally easy to use.

Frequent meetings took place between the team leader and the computer programmer to work out the possibility of having the required features and functions within the existing resources, expertise and time constraints. For example, we could not include some features either because there was insufficient time to work on them or we needed more time to explore them in order to gain knowledge and expertise. We came up with satisfactory alternatives and planned for better features in the future.

The language staff member uploaded the test items according to their level, skills tested, and type of activity, using the Item Editor tool, thus building the test item pool, and its sub-test item pools. This was a time-consuming activity and needed great attention.

Item selection Algorithm

The next step was to work out a test item randomization system to produce pen-and-paper (P & P) unique tests to field- test the NEPTON test. The two language faculties worked together to develop their own algorithm. The aim was twofold: develop an algorithm which would aim at the most comprehensive item selection as possible (language level, type of activity, type of skill, type of text, number of items), and a system which would be within our human, monetary and software constraints. The result was an algorithm which generates unique tests through test (Test 1 or Test 2) and random choice of test items. Each test had questions at 6 levels. Each test consisted of two slides per language level. Each slide contained 9 items, randomly chosen from sub-pools of different item categories: structure-sentence, vocabulary sentence, structure text, vocabulary text, sign, and reading comprehension text. The total was 18 items per level by 6 levels, totalling 108 items for each P & P field test paper for each student.

Table 1: Item Selection Algorithm

	TEST 1		TEST 2	
BENG-50 & BENG-80	**Slide 1**	**Slide 2**	**Slide 1**	**Slide 2**
	SS 5 VT 4	VS 3 ST 5 S 1	SS 5 RC 4	VS 3 ST 5 S 1
	9 items	**9 items**	**9 items**	**9 items**
	18 items per level		**18 items per level**	
BENG-90; BENG-100; ENG-100 & ENG-101	**Slide 1**	**Slide 2**	**Slide 1**	**Slide 2**
	SS 5 RC 4	VS 4 ST 5	SS 4 VT 5	VS 5 RC 4
	9 items	**9 items**	**9 items**	**9 items**
	18 items per level		**18 items per level**	
TOTAL	18 items by 6 levels=108 items		18 items by 6 levels=108 items	

Although we knew that the test was too long, we had to field test it in this form for three reasons:

(a) to have each item used by as many students as possible for better item analysis

(b) to expose the items to as many students and staff as possible for feedback

(c) to test the item selection algorithm

NEPTON Pen-and-Paper Field-Testing

The NEPTON test was initially randomly generated using the above algorithm, and field-tested in pen-and-paper form in May 2004. This helped improve the content of the test, and the item selection algorithm, and formulate the cut off points.

Due to the urgency to implement the test in the new academic year, we field-tested it during the end of semester 2, just before the exams and during exam preparation time. That was not the best of times to ask both staff and students to field-test it. As a result, some staff chose not to participate in the field testing with their students. Others did not invigilate properly. Some students did not take it seriously. The team had to organise, monitor the field testing and process its data while dealing with final exam writing and correction, and final mark submission.

There were also some IT related problems: due to time and expertise constraints, there were delays in the preparation of the electronic test. This resulted in delaying the test item uploading process and consequently the field-testing. The agreed slide design was not consistently followed during the randomized test item selection, and as a result, the programming of the algorithm needed to be checked again by the computer programmer. Scanning of the answer-sheets had to wait because of other college administrative scanning priorities. This delayed the data analysis and the preparation for the electronic test trial.

The pressure on the team members during the field testing was immense. Above their usual duties, they had to implement the field testing at three campuses in three different cities. The computer programmer had to incorporate the test item selection into the system, generate enough number of unique tests and print them out, prepare the answer-sheets and manage their scanning. The two language faculty had to organise the field testing in the three campuses: invigilation, test distribution and collection, provision of pens, rubbers, and envelopes, instructions to the invigilators, test collection and collation, thorough check of the algorithm, scanning system and test results. There were certainly some sleepless nights then, since the work was more than the available time human resource. All that work had to be done at no additional costs.

Another constraint after the field testing was the lack of sufficient staff to conduct the data analysis. The test was checked again by a group of four volunteer, native speaker professionals, with experience in testing. The language staff member processed the instructors' and invigilators' input, the test-takers' post-test open-ended questions, the test team observations, and the extra moderators' input, and made the necessary improvements. As a result, many test items were revised and the test item pool of the original number of 1500 items went down to 1084 items:

Table 2: Test Item Bank and Sub-Pools

ITEMS							
Level	SB-S	TB-S	SB-V	TB-V	SB-RC	TB-RC	
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
ENG-100	46	25	55	24	0	16	166
ENG-101	33	25	27	20	0	16	121
TOTAL	280	220	175	196	41	172	1084

These discrete items were level-based and of the following activity types: there were 280 items in the Sentence-based Structure (SB-S) sub-pool, 220 items in the Text-based Structure (TB-S) sub-pool, 175 items in the Sentence-based Vocabulary (SB-V) sub-pool, 196 items in the Text-based Vocabulary (TB-V) sub-pool, 41 items in the Sign-based Reading Comprehension (SB-RC) sub-pool, and 172 test items in the text-based Reading Comprehension (TB-RC) sub-pool in all levels. At the same time, there were 211 items of all activity types in the BENG-50 sub-pool, 244 items in the BENG-80 sub-pool, 212 items in the BENG-90 sub-pool, 130 items in the BENG-100 sub-pool, 166 items in the ENG-100 sub-pool, and 121 items in the ENG-101 sub-pool, a total of 1084 items in the whole item bank. The following are examples of the test activity types and skills tested, as they appear in the online testing environment:

Figure 4: Interface 1, Sentence-Based Activity, Testing Grammar or Vocabulary.

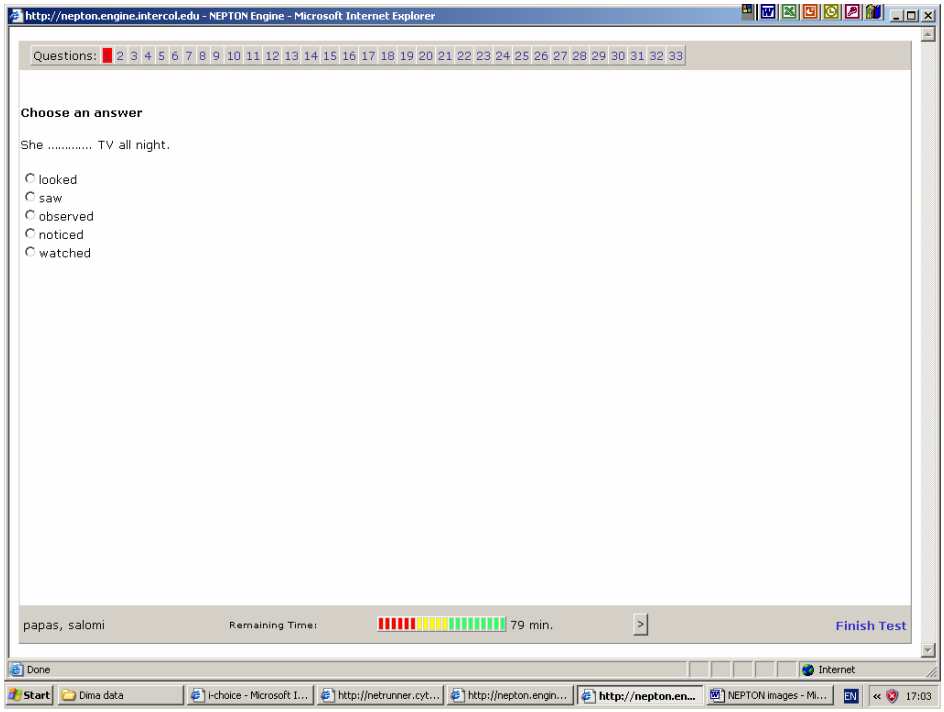


Figure 5: Interface 2, Text-Based Dropdown Activity Type, Assessing Grammar or Vocabulary.

Figure 6: Interface 3, Sign-Based Activity Type, Assessing Reading Comprehension with Visuals.

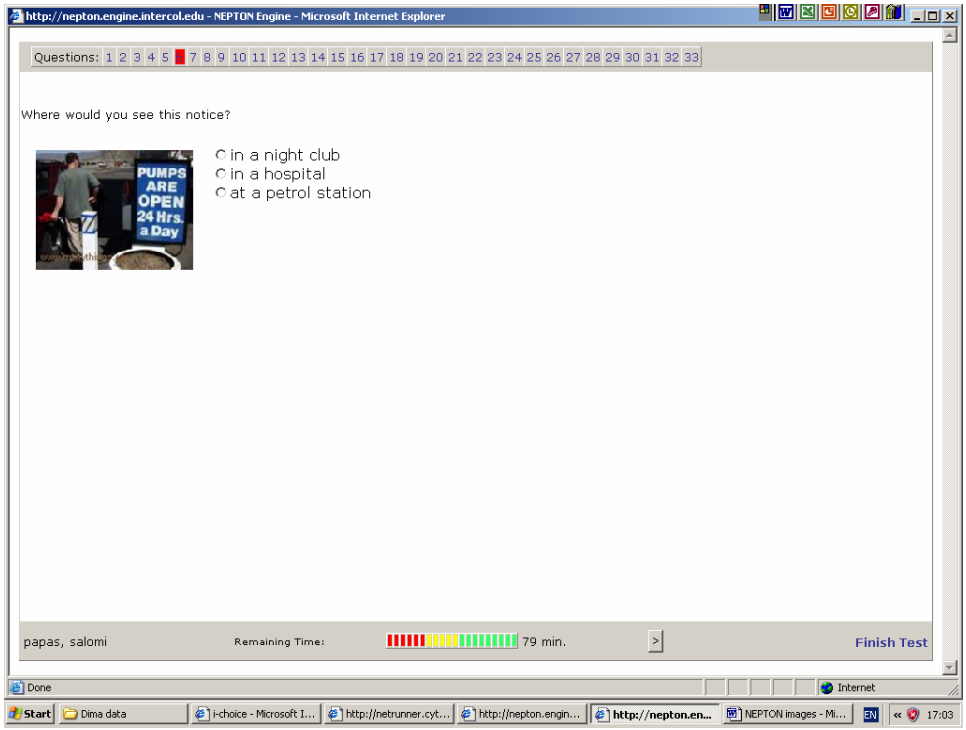
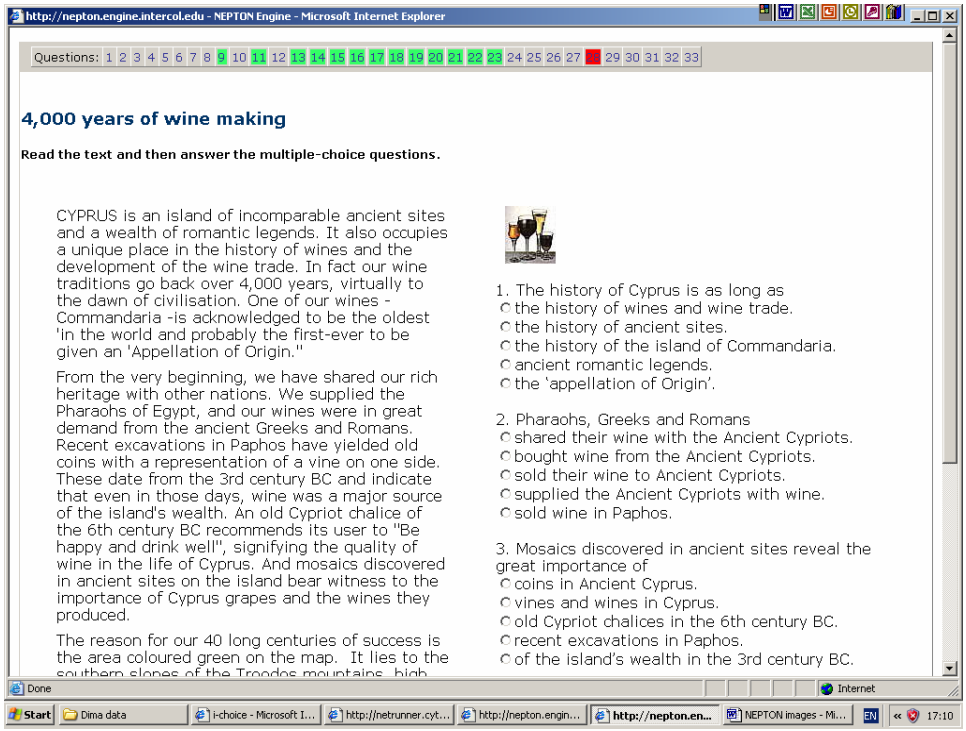


Figure 7: Interface 4, Text-Based Multiple-Choice Activity Type, Assessing Reading Comprehension



Revised NEPTON Slide Paradigm

The two language faculties spent a lot of time and finally worked out a new slide algorithm. There was a lot of discussion on the selection criteria of items to be included for each level slide. As a result the test was shortened from 65 questions (108 items) to 33 questions (54 items). This new slide paradigm was used to generate the final version of the NEPTON test, in both its electronic and pen-and-paper form.

Table 3: Revised, Final NEPTON Slide Paradigm

	TEST 1	TEST 2
Slide 1: BENG 50		
9 items per slide	SB-S: 3 TB-S: 5 SB-RC: 1	SB-V: 4 TB-V or TB-RC: 4 SB-RC: 1
Slide 2:BENG 80		
9 items per slide	SB-V: 4 TB-V or TB-RC: 4 SB-RC: 1	SB-S: 3 TB-S: 5 SB-RC: 1
Slide 3:BENG 90		
9 items per slide	SB-S: 4 TB-S: 5	SB-V: 5 TB-V or TB-RC: 4
Slide 4:BENG 100		
9 items per slide	SB-V: 5 TB-V or TB-RC: 4	SB-S: 4 TB-S: 5
Slide 5:ENG 100		
9 items per slide	SB-S: 4	SB-V: 5

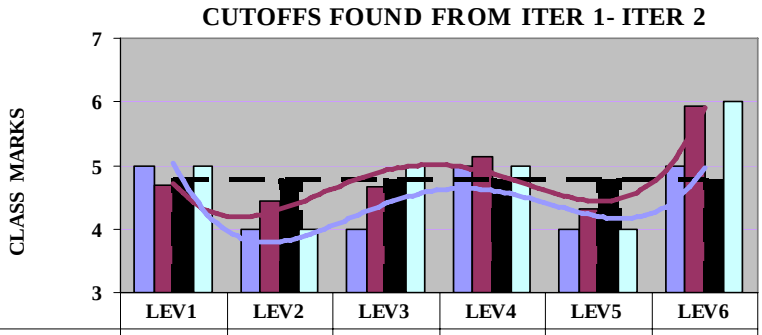
		TB-S: 5		TB-V or TB-RC: 4
Slide 6:ENG 101				
9 items per slide		SB-V: 5 TB-V or TB-RC: 4		SB-S: 4 TB-S: 5
9 items x 6 slides		54 items (33 questions)		**54 items (33 questions)**
Total test length= 9 items x 6 levels = 54 items				
Activity types: Sentence-based Structure (SB-S), Text-based Structure (TB-S), Sentence-based Vocabulary (SB-V), Text-based Vocabulary (TB-V), Sign-based Reading Comprehension (SB-RC), & Text-based Reading Comprehension (TB-RC)				

Either Test 1 or Test 2 is randomly chosen. For each test, each test-taker is then presented with a series of six slides, not two as for the field testing, but one slide per level (BENG-50, BENG-80, BENG-90, BENG-100, ENG-100 and ENG-101). Each slide includes nine items. These items are chosen from different sub-pools of items, representing different types of activities and different skills (Table 5, Test 1 column 2 and Test 2, Column 3). Test-takers answer thirty-three randomly chosen questions, including fifty-four items in all.

Cut-Off Points

Analysis of the results of the pen-and-paper field testing did not only help to improve the test, the test items, and the item selection algorithm, but also provided the method which helped generate the NEPTON cut-off points that were necessary for the automatic placement of test-takers. At our college, we do not have an expert in language testing statistics. The two language faculties took into consideration the hybrid nature of the test and available human and software resources and came up with a system to calculate cut off points for the test. The second language faculty and a volunteer statistician ran a statistical analysis of the responses

to the pen-and-paper field testing version results. These yielded the parameters that were used by the algorithm to interpret meaningfully the NEPTON scores against the college English Language Programme Competence levels, which reflect the syllabus framework, the teaching materials and the assessment procedures.



	LEV1	LEV2	LEV3	LEV4	LEV5	LEV6
ITER-2 CUTOFFS FROM ITER-1	5	4	4	5	4	5
CLASS AVERAGES	4.70	4.43	4.66	5.13	4.31	5.94
MASS AVERAGE	4.76	4.76	4.76	4.76	4.76	4.76
FOUND CUTOFFS FOR ITER-3	5	4	5	5	4	6

As the administration wanted to proceed with the test quickly, the whole process had to be completed during the summer holiday period, through on-going communication and cooperation with the project leader, the head of the language department, and the administration.

There were still many things to do after working out the cut off points: during the iteration process, the project leader prepared an Invigilator's instruction booklet, and designed and wrote the content of the electronic tutorial and the trial test, which aimed to familiarize test-takers with the test and the basic computer skills needed for the test. These had to be done in agreement with the rest of the team. The computer programmer then turned the electronic tutorial and trial test into their electronic form and linked it to the test. In addition, the language faculty prepared a sample test in printed form to be given to incoming students for test familiarization. Moreover, the two language faculties prepared hand-written task

marking criteria and ran marker training sessions. We also had to organize the test trial for September.

Test Trial

The test trial took place in early September 2004. There were about 800 test-takers involved in the autumn NEPTON administration. The whole team cooperated with the Orientation Week administrator to organize it. The test was hosted on the college server and delivered via the Internet. Eight computer labs were used, a total of 127 computers at a time. The test-takers were invigilated in the computer laboratories at the college, by trained college administrators who helped test-takers through the tutorial (test and basic computer skills familiarisation session). The project team and two lab assistants were on stand-by for all laboratories at all times.

Data Analysis for Test Reliability and Validity

The next section describes the data analysis, which helped test the NEPTON's reliability and validity and the constraints encountered during this process.

Item Analysis

The item analysis was one of the most challenging tasks for the project leader. Not only because she had never done this before, but also because the hybrid nature of the test required an item analysis system that could not be found in existing literature and had to be worked out. Thorough examination of existing item analysis systems (Brown, 2003; Special Connections, Item Analysis 2005; Scoring Office, Michigan State University 'Item Analysis' website 2005; *Test Scoring Statistics Guide* 2005; Alderson et al., 1955; Test Scoring Statistics Guide, Interpreting the Reports 2005), and the test's hybrid nature, and a lot of hard

thinking, resulted in a system which made the item analysis for our test possible. She decided that an acceptable facility value for each item would be between 25% and 80%. To establish the facility value of the items, each item was ranked from high to low level. The total number of students 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 students who took the item to establish the item's facility value. 61% of the test items were found to have good facility value.

The NEPTON Discrimination Index (D.I.)

What is considered as the ideal discrimination index (a positive DI above 0.30) by the *Test Scoring Statistics Guide, Interpreting the Reports* (2006) was used to calculate the discrimination index of each NEPTON test item. To establish this, all 1084 test items were ranked from high to low level. The total number of students who took each item was recorded next to each item, in two categories (high and low score groups). The number of correct answers (CH) in the high score group and the total number of students 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 (CL) in the low score group and the total number of students who took each item at low score group (SL) were recorded. The DI was calculated for the low score group as well (DL). The low and high difficulty indices were then calculated to arrive at the DI of each item. Four hundred and thirteen out of 1084 test items had an acceptable DI above .30. About 40% of the test items were found to have acceptable Discrimination Index.

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 process helped keep in the test only those items that were well centered

and discriminated well between the high and the low scoring students. The rest of the test items were reviewed at a later stage. As a result about 60% of the items were considered good items to be included in the test. Improvement of the existing items then took place and new items were being developed.

The NEPTON inter-item consistency: split-half reliability index: NEPTON test reliability was estimated measuring the *inter-item consistency* (Alderson et al., 1995). Although she was helped by another statistician, it was hard to find time to work together on the data analysis. Despite this, they managed to simulate the parallel forms method by calculating the *split-half reliability* index. This involved dividing a 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. A perfectly reliable test would have a reliability index of +1.0. As the table below indicates, the two halves of the NEPTON test correlated strongly, thus suggesting a high NEPTON test reliability.

Table 7: NEPTON Correlations

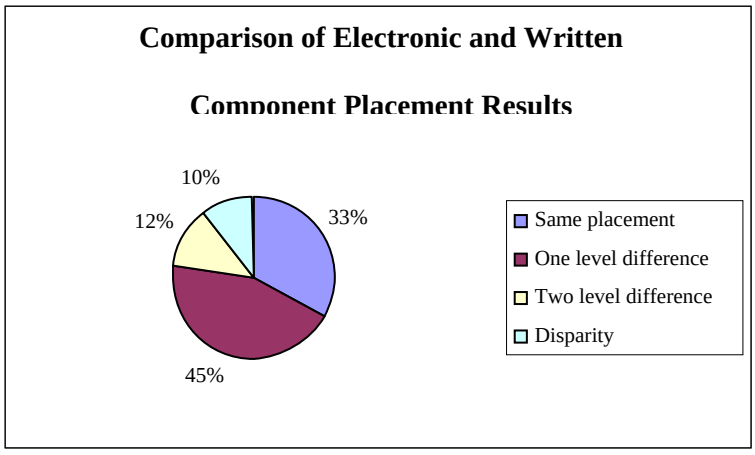
		Odd number	Even number
Odd number	Pearson Correlation	1	.822(**)
	N	866	866
Even number	Pearson Correlation	.822(**)	1
	N	866	866

** Correlation is significant at the 0.01 level (2-tailed).

Other Test Reliability Measurement Techniques

The project leader made a conscious effort to ensure the effects of sources of reliability as indicated by Hughes (1989), and Dunkel (1999) were minimised. This was done through the following means: large item bank; moderation; discrete items and objective scoring; detailed test specifications, which include cut-off points; clear computer interface; care for general and individual factors such as familiarity with the test format (sample pen-and-paper test, electronic tutorial and test trial, test orientation); care for situational factors such as test administration conditions. Other factors taken into consideration were: the construction process of the item bank was based on the needs analysis, English language programme curriculum, and Test Specification. We also had the pen-and-paper cut-off points as a starting point. We studied the gross proportion of statistics which indicated from previous years the proportional percentage of students allocated to each level and there was no indication there were differences in this year's intake. We compared the electronic placement with the written component placement results to find out how they correlate. The figure below indicates how they compare:

Figure 8 Comparison of electronic and written components of NEPTON results



33% of both results were exactly the same. Although a total of 57% (45% one level and 33% two level difference) indicates a good correlation between the two components of the test, a higher correlation of the same placement would have been more satisfactory. The 22% (two level difference and disparity) was of concern and needed investigation and improvement.

Test Validity

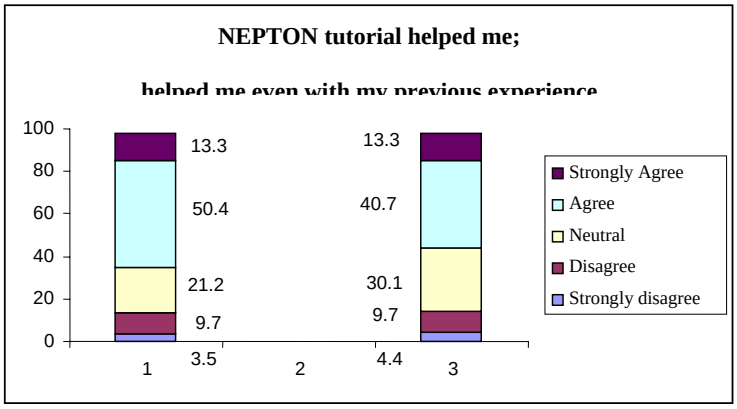
The sample used in the test trial was of adequate size for the test validation. There were more than 800 test-takers. About 260 answered the pre-test questionnaire and about 131 answered the post-test questionnaire. The sample is representative of the population for which the test is intended in age, experience and background. The language levels of the test provide an adequate basis for validating the instrument. The large size of the item pool (1084 items) also secured higher test validity.

Internal Face Validity

Test-takers had a choice of taking the test electronically or in pen-and-paper (P & P) format. At the beginning, out of more than 800 students, only five students said they wanted to do the (P & P) NEPTON but when they did the NEPTON tutorial, two of them changed their mind and did NEPTON instead, and only three opted for the (P & P) NEPTON option. This suggests that the majority of test-takers, preferred to take the placement test electronically.

The language faculty developed pre and post NEPTON test questionnaires. These were checked by the second statistician, who also helped the language faculty come up with the necessary techniques for the data analysis. These Pre and Post test-taker's NEPTON test questionnaires indicated satisfactory comfort of test-takers with test. Test-takers were also asked whether or not the tutorial has helped them.

Figure 9: The NEPTON Tutorial and Computer Familiarity



The student NEPTON post-test data indicates that the majority (13.3% strongly agreed and 50.4% agreed) that the tutorial was helpful to them. Even those with some previous computer experience strongly agreed (13.3%) or agreed (40.7%).

Table 8: NEPTON Post-Test Questionnaire – Skill Sections

	SD	D	N	A	SA
I found the structure/vocabulary (sentence-based) section manageable.	1.8%	10.6%	21.2%	44.2%	21.2%
I found the structure/vocabulary (text and dropdown menu selection) section manageable.	1.8%	3.5%	27.4%	44.2%	20.4%
I found the reading					

comprehension sign section manageable.	1.8%	7.1%	23.0%	43.4%	20/4%
I found the reading comprehension (text and multiple-choice) section manageable	0.9%	8.0%	19.5%	49.6%	17.7%

Key: SD: strongly disagree; D: disagree; N: neutral; A: agree; SA: strongly agree

Based on the data above, most test-takers found the different types of questions covering the different skills manageable. The majority also found the instructions clear (43% agree and 46% strongly agree).

Table 9: NEPTON Post-Test Questionnaire – Topics and Variety of Activities and Test Length

		SD	D	N	A	SA
1	Interesting topics	2.7%	8.0%	22.1%	38.9%	23.0 %
2	Enough variety of activities	1.8%	9.7%	23.9%	46.0%	15.9 %
3	Appropriate test length.	1.8%	7.1%	26.5%	38.1%	24.8 %

Key: SD: strongly disagree; D: disagree; N: neutral; A: agree; SA: strongly agree

The test-takers also found the NEPTON topics of the electronic part of the test interesting (38% agree and 23% strongly agree), while 22.1% were neutral, 8% disagreed and 2.7% strongly disagreed. The test takers also found that there was sufficient variety of activities (46% agree and 15.9% strongly agree), while 23.9% were neutral, 9.7% disagreed and 1.8% strongly disagreed. Test-takers seem to be happy with the length of the electronic (26.5% neutral 38.1% agreeing and 24.8% strongly agreeing) test. Again, the proportion of neutral responses is substantial and may need consideration.

The results above as a whole 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 test, test items and test components.

Internal, content validity

At our college there are 24 full-time and 23 part-time practising English lecturers across all three campuses. Six of them (two from each campus) with expertise in English Placement Testing and development were asked to examine NEPTON's content validity by doing the following:

(a) Study the NEPTON Test Specifications

(b) Study two sample Pen-and-Paper tests

(c) Study NEPTON

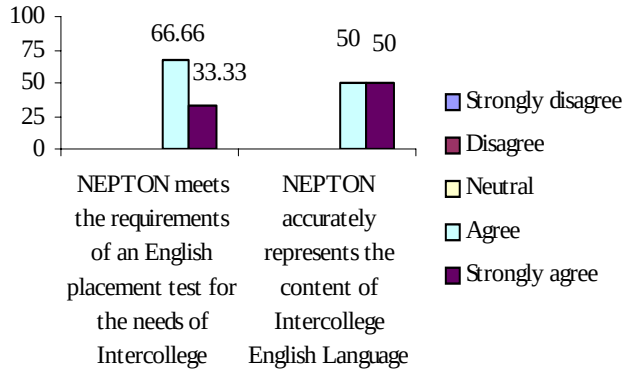
(d) Compare all three tests (The NEPTON and two sample pen-and-paper tests) with the Test Specification by rating the test on a questionnaire (prepared by the project leader) according to the degree to which it met certain criteria

The questionnaires were developed by the project leader, and checked by the second volunteer statistician, who again helped the project leader come up with the necessary techniques needed to analyse these data. This again was time consuming because the people worked in three campuses in three different cities, so coordinating the whole process involved

travelling. By the end of this data analysis the project leader became quite an expert in the use of Excel (2003) and SPSS (2000) programmes. However, it was learned the hard way, within impossible and frustrating time constraints. The following section provides a selection of data relating to the internal content validity.

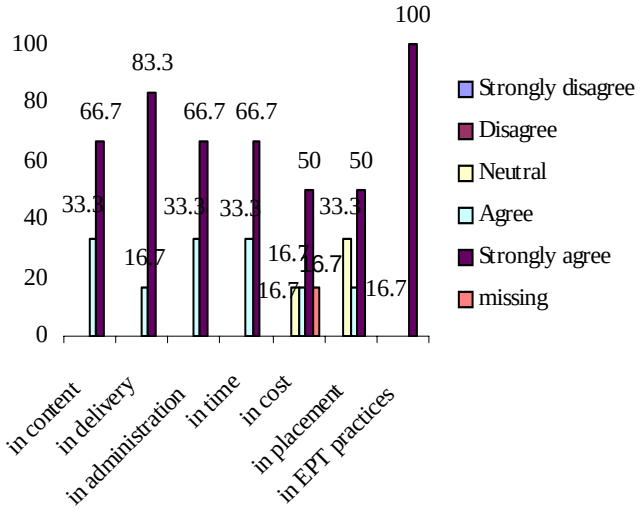
Experts' Evaluation

Figure 10: Experts NEPTON Content Validity Evaluation



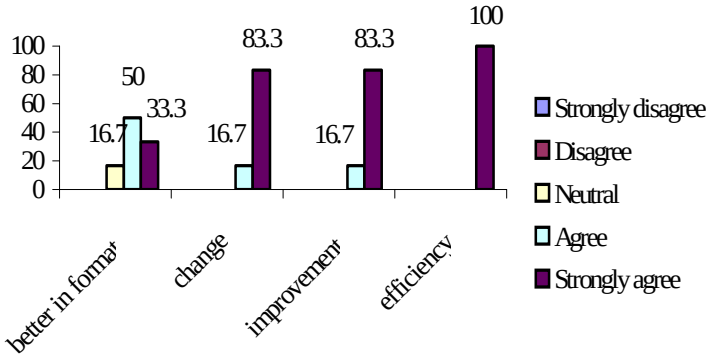
The data in Figure 10 above show that 33.33% of the respondents strongly agreed and 66.66% agreed that the NEPTON test met the requirements of an English placement test for the needs of our college students. 50% strongly agreed and the other 50% agreed that the NEPTON accurately represented the content of the college English language programme at the various levels.

Figure11: NEPTON Efficiency



From figure 11, it is evident that 66.67 strongly agreed that the NEPTON is more efficient in content, administration and time. 83.3% strongly agreed and 16.7% agreed that the NEPTON is more efficient in delivery mode. 50% strongly agreed that it is more efficient in cost and placement. 16.7 agreed 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 or not NEPTON was more efficient in placement. All respondents strongly agreed that the NEPTON test has brought improvement in EPT practices.

Figure 12: NEPTON Better than Previous EPT; it has brought change, improvement and efficiency



As indicated in the data above, 16.7% felt neutral, 50% agreed and 33.33% strongly agreed that NEPTON was better in format than the previous EPT. Finally, according to illustration 3, 83.3% of the respondents strongly agreed and 16.7% agreed that NEPTON had brought change, and improvement and all agreed that it brought improved efficiency to our college EPT practices.

Limitations

One of the test's limitations was that there was not enough opportunity for all items to be tested adequately by a great number of test takers before it was first used. This is becoming more possible with the on-going application of the test and the availability of more data for such item analysis. More time was needed for all team members to work on the project. More time was also needed to adequately inform and involve all stakeholders on the concept of the new test.

Conclusions

The description of the human and other resources of this project clearly indicates the restricted resources available for the implementation of the NEPTON test and how

challenging it has been to combine current theories and practices in language computer-based testing with the practical realities of an institution with limited resources. During the whole process, we felt the need for more people who could have helped in the following areas: content, programming, test item writing, moderation, field testing and marking, and data analysis. It was clear that we needed more expertise beyond the language faculty experts in conducting statistical analysis; in combination, all phases of the project needed more time to be thoroughly dealt with. In addition, with more time, the project team and other faculty would have had the opportunity to extend their expertise even more. More test fine tuning would have been possible, if deadlines were not so strict. If more time and financial allocation had been available, more staff would have been encouraged to contribute and benefit from this project. It was also clear that greater investment in hardware and software would have eliminated or solved more quickly, some of the problems we encountered.

However, we managed to deliver a good test on time, and within our local human and resource constraints. After two years of use and improvement, keeping within our resources and utilizing no additional time or monetary allowances, we feel we have managed to develop and implement a New English Placement Test Online with success and which offers Computer assisted language testing some innovative features. It has been hard work but very useful and extremely rewarding.

References

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

A.L.T.E. (2002). *Common European framework of reference for languages: Learning, teaching, assessment, language examining and test development*. Strasbourg: Language Policy Division

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

Burston, J., & Monville-Burston, M. (1995). Practical design and implementation considerations of a computer adaptive foreign language test: The Monash/Melbourne French Cat. *CALICO Journal, 13*(1).

Brown, J. D. (2003) Norm-referenced item analysis (item facility and item discrimination), *Shiken: JALT Testing & Evaluation SIG Newsletter, 7*(2), 16-19), [Online]. Available: http://www.jalt.org/test/bro_18.htm (Accessed 14 February 2005).

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology, 1*(1), 44-59. http://llt.msu.edu/vol1num1/brown/default.html (Accessed 19 June 2002).

Brown, J. D. (1995). *The elements of language curriculum: A systematic approach to program development.* New York: Heinle & Heinle.

CB IELTS (2005). [Online]. Available: http://www.ieltsindia.com/cb_ielts.htm (Accessed 03 March 2007).

Chalhoub-Deville, M. (2001). Language testing and technology: Past and future. *Language Learning Technology, 5*(2), 95-98, [Online]. Available: http://ll.msu.edu/vol5num2/emerging/default/html (Accessed 19 June 2002).

College Placement Test: Marking Guidelines and Sample Placement

College English Syllabi*: BENG 50, BENG 80, BENG 90, BENG 100, ENGL 100, ENGL 101*

College Course Outlines*: BENG 50, BENG 80, BENG 90, BENG 100, ENGL 100, ENGL 101*

College previous English Placement Tes*t*, pen-and-paper format

College Prospectus, 2003-2004

Council of Europe. (2001). *Common European framework of reference for languages, learning, teaching and assessment.* Cambridge.

Dunkel, P. (Ed.). (1991). *Computer-assisted language learning and testing*. New York: Newbury House.

Dunkel, P. (1997). Computer-adaptive testing of listening comprehension: A blueprint for CAT development. *The Language Teacher Online*. [Online]. Available: http://lange.hyper.chubu.ac.jp/jalt/pub/tlt/97/oct/dunkel.html (Accessed 19 June 2002).

Dunkel, P. (1999). Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning & Technology, 2*(2) 77-93. [Online]. Available: http://www.multingles.net/docs/tai_dunkel.htm (Accessed 22 July 2003).

Excel (2003). Microsoft.

Fulcher, G. (2001). *Resources in language testing page.* [Online]. Available: http://www.surrey.ac.uk/ELI/ltr.html. (Accessed April 1, 2001).

Fulcher, G. (2000). Computers in language testing. In P. Nrett, & G. Motteram (Eds.), *A special interest in computers* (pp. 93-107). Manchester: IATEFL. [Online]. Available: http://www.dundee.ac.uk/languagestudies/ltest/ltrfile/Computers.html. (Accessed 26 September 2003).

Godwin-Jones, B. (2001). Language testing tools and technologies. *Language Learning Technology, 5*(2), 8-12. [Online]. Available: http://llt.msu.edu/vol5num2/emerging/default.html (Accessed 19 June 2002).

Half-baked. (2004). *Hot Potatoes*. Version 6.0.3. Half-baked Software, inc. [Online]. Available: http://web.uvic.ca/hrd/halfbaked/ (Accessed 29 December 2003).

Heaton J. B. (1988). *Writing English language tests.* UK: Longman.

Henning, G. (1987). *A guide to language testing: Development, evaluation, research.* New York: Newbury House.

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

Kitao, S., & Kitao, K. (2002). *Test Specifications*. TESL-L Electronic Discussion Forum for teachers of English as a second or foreign language, [Online]. Available: http://ilc2.doshisha.ac.jp/users/kkitao/library/article/test/design2.htm#specification (Accessed 23 October 2003).

Kitao, S., & Kitao, K. (1996). Testing communicative competence. *The Internet TESL Journal*, 2(5), [Online]. Available: http://iteslj.org/Articles/Kitao-Testing.html (Accessed 26 September 2003).

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

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

Microsoft Office Excel, (2003)

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

Papadima-Sophocleous, S., & Apraksin, D. (2005). *New English placement test online*. Cyprus.

Papadima-Sophocleous, S. et al. (2005). *The COBLE (committee of BENG level English) Report, Review of the College English Programme.* Draft. Cyprus.

Quia. (1998-2006). Quia Corporation. [Online]. Available: http://www.quia.com/ (Accessed 16 February 2000).

Quick Placement Test (2001). UK. University of Cambridge, Oxford University Press. Windows edition CD-ROM.

Question Mark Perception Question Mark Computing Ltd. (n.d.). Windows edition CD-ROM sampler.

Roever, C. (2001). Web-based language testing. *Language Learning Technology*, *5*(2), 84 [Online]. Available: http://llt.msu.edu/vol5num2/emerging/default.html (Accessed 19 June 2002).

Sceppa, D. (2002). *Microsoft ADO.NET* (Core Reference) (Hardcover) USA: Microsoft Press, Redmond Washington.

Scoring Office, Michigan State University 'Item Analysis' (2005)

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

Special Connections, Item Analysis (2003) [Online]. Available: http://www.specialconnections.ku.edu/cgi-bin/cgiwrap/specconn/main.php?cat=assessment&section=main&subsection=qualitytest/item (Accessed 14 February 2005).

SPSS (2000)

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.

Test Scoring Statistics Guide, Interpreting the Reports, [Online]. Available: http://helpdesk.kent.edu/howto/testscoring/stat/. (Accessed 14 February 2005).

TOEFL Testing Program. [Online]. Available: http://www.toefl.org/tstgprog01.cfm (Accessed 31 March 2001).

TOEFL CBT. (1998). Available: http://www.toefl.org/tstgprog01.cfm (Accessed 3 March 2007).

TOEFL iBT. (2005). Available: http://www.toefl.org/tstgprog01.cfm (Accessed 3 March 2007).

Walther, S. (2003). *ASP.NET Unleashed*. (2nd ed.). USA: SAMS Publishing.

WebCT. (2004). Version 3.0. WebCT, inc. [Online]. Available: http://www.webct.com (Accessed 08 September 2001).

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

Weir, C. J. (1995). *Understanding and developing language tests.* Great Britain: Phoenix ELT.

West, R. (1994). Needs analysis in language teaching. *Language Teaching, 27*(1), 1-19.

Witkin, B. R., & Altschuld, J. W. (1995). *Planning and conducting needs assessments: A Practical guide.* London: SAGE.