Developing Computerized Language Proficiency Tests at NIE

Victoria Y. Hsui, Anthony Seow & Chew Lee Chin
Nanyang Technological University
National Institute of Education

Introduction

This article discusses work-in-progress concerning a research and development project at the National Institute of Education to develop a computerized English language test. The test, which aims at addressing the various needs for language testing at the institution, is a comprehensive one. It will assess all the main components of language, including speaking, listening, grammar, vocabulary, reading, and writing. The computerized test, called NIE-CELT (NIE-Computerized English Language Test) is a battery of tests comprising a range of discrete-point and integrative-skills items (Oller, 1973) which are complementarily targeted at measuring the candidates' proficiency in the English Language. The various language components that are tested are organized into 7 Domains:

1 Reading aloud and Speaking
2 Listening comprehension
3 Grammar
4 Vocabulary
5 Reading comprehension
6 Cloze (assessing reading comprehension, vocabulary and grammar in context, cohesive devices, etc.)
7 Essay writing

Need for a Computerized Test

Conducting English language testing at the institution has been largely a complex and difficult task because of the many proficiency tests developed and administered by various programs and for various purposes. Some of these tests are taken by large numbers of candidates (up to 1,000) and are conducted several times a year. The tests are currently administered in the paper-and-pencil format. For tests taken by large numbers of candidates, each time they are administered, at least 35 academic staff members need to be involved in the invigilation of the tests, and in the marking and entering of marks. A number of administrative staff are also involved in the preparation of test papers and test paraphernalia. Technology is therefore explored to see how it can help to ease these burdens. The NIE-CELT is the result of this exploration.

Progress Report

To date, test items for Domains 3 to 6, that is, Grammar, Vocabulary, Reading Comprehension, and Cloze have been developed and pilot-tested through two parallel test "forms": Form A and Form B. Questions for Domains 1, 2 , and 7 (Reading aloud and Speaking, Listening comprehension, and Essay Writing) are in the process of being developed. The questions that have been developed are the result of the collaborative effort of a testing team, which comprises eight members of the English Language and Applied Linguistics Division of the National Institute of Education.

For Domains 3 to 6, which are captured in two parallel tests, Form A and Form B, test items were constructed so that items appearing in one Form also appears in the other Form. These parallel items reflect each other in format, style, and language elements, but are different in content. For the pilot tests, the two Forms were administered in the paper-and-pencil format. The test items were, however, constructed with technology in mind – to utilize computer technology to test large numbers of candidates in ways that are not possible with the paper-and-pencil format, where test results need to be compiled quickly and expeditiously. Technology will be exploited as much as possible in administering and scoring the test items which include open-ended questions, word unscrambling, "point, drag, and drop" items, and multiple-choice questions.

Forms A and B were pilot tested with NIE students from three programs: Diploma in Education, BA/BSc, and Postgraduate Diploma in Education. Students were randomly assigned to take either one of the Forms, under supervised conditions. They were allowed two hours to take the test. The tests were scored manually, with the scorer instructed to look out for and record plausible answers that were not already addressed in the marking scheme. The test scores were then subjected to psychometric analyses. All these steps were taken to provide the researchers input to guide the computerization of the tests.

The following sections will discuss the analyses of the tests (Form A and Form B) and the beginnings in the development of a prototype for the computerization of the test.

Test and Item Analyses

Two parallel test forms, sampling language content for domains 3, 4, 5 and 6, were constructed based on the test specifications as shown in Table 1. A total of 123 test items were used in each test form, with the specified number of items for each domain.

Table 1: Test Specifications

Domain	Content	No. of Test Items
3	Grammar	34
4	Vocabulary	30
5	Reading Comprehension	12
6	Cloze	47
		Total: 123

Table 2: Summary Statistics by Test Form

Statistics	Form A	Form B
No. of items	123	123
No. of examinees	72	69
Mean Score	85	80
Std. Deviation	14.8	14.1
Skew	-.7	-.3
Kurtosis	1.6	-1.0
Minimum	28	49
Maximum	112	104
Median	84	83
Alpha	.92	.90
Std Error Measurement	4.26	4.47
Mean Proportion Correct	.69	.65
Mean Item-Total	.31	.27

The two data sets were analyzed using ITEMAN, an item and test analysis computer program (Assessment Systems Corporation, 1994). The item statistics obtained provided useful information on the difficulty and discrimination of each test item. Table 2 shows the summary statistics. Altogether 72 students took Form A and 69 took Form B. Compared to Form A, a slightly lower mean test score was obtained for Form B (80 versus 85). Results show good alpha coefficients ranging from .90 to .92 for the two test forms, and these attest to good test reliability. From the descriptive statistics, the results indicate satisfactory overall comparability of the two test forms that were constructed based on the same test specifications.

Test comparability of the two parallel forms was further analyzed. Chart 1 shows the mean scores by domain. Except for Domain 3, higher mean test scores were found on Domains 4, 5 and 6 of Form A. But these differences in mean scores were not significant, and the results attest to the comparability of the two test forms.

Chart 1: Mean Test Scores by Domain

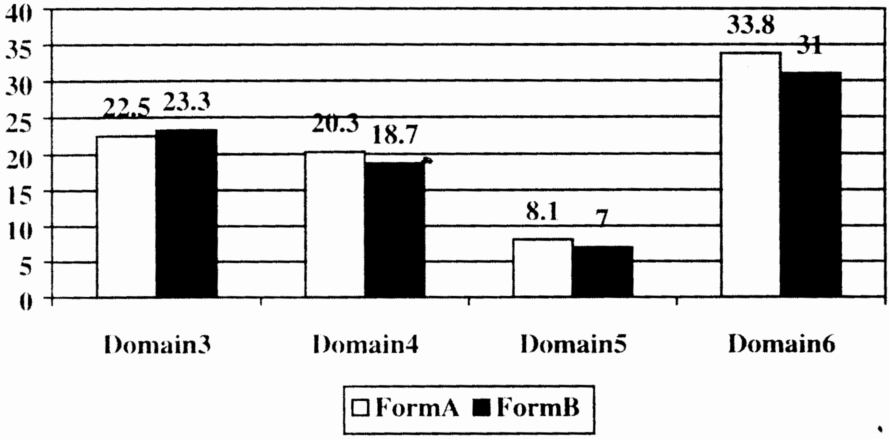


Chart 2: Mean Test Scores by Program

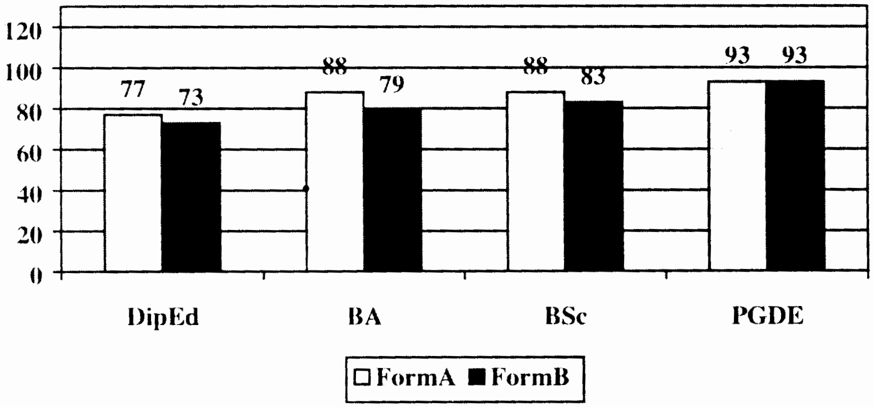
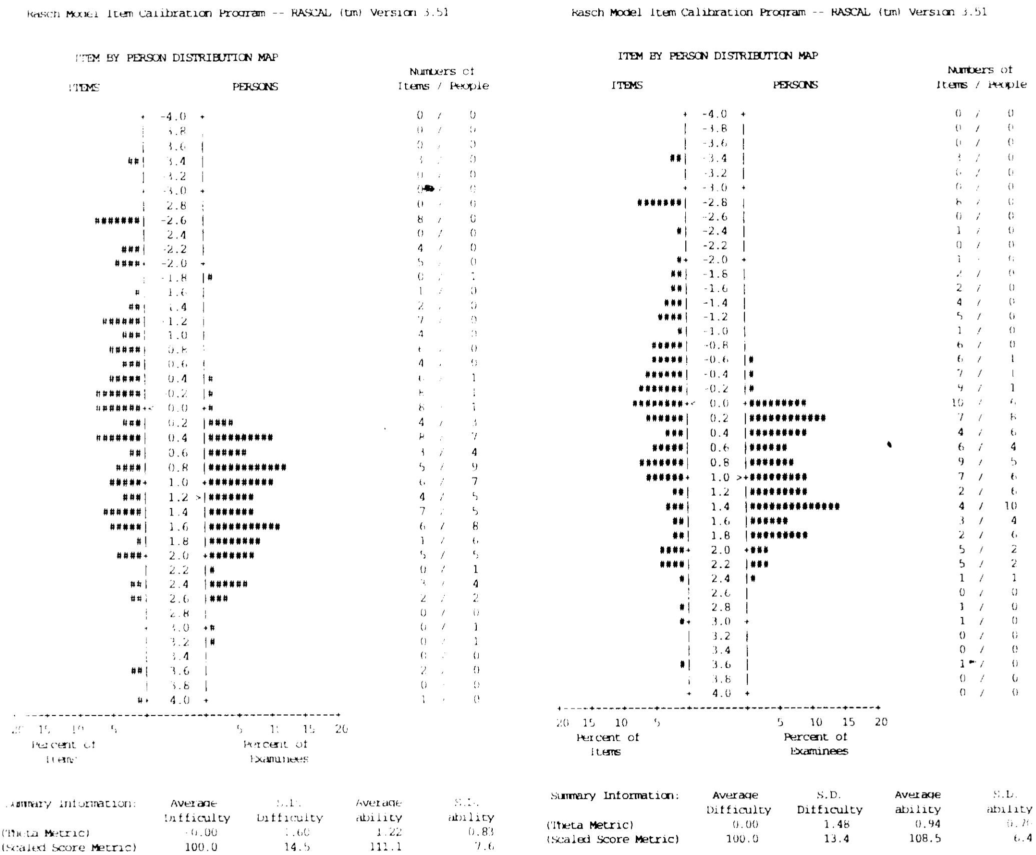


Chart 2 shows the mean test scores by program. There was comparable performance by students in each program, namely, DipEd, BA/BSc and PGDE on each test form. But the PGDE students outperformed the DipEd and the BA/BSc students, with the DipEd students having the lowest mean test scores. These results are expected given that the students in the PGDE program have had university education and are, therefore, cognitively and intellectually more advanced than the other three sub-groups of students.

Using the RASCAL program (Assessment Systems Corporation, 1994), the two data sets were subjected to a preliminary item calibration based on the 1-parameter logistic model. Compared to classical psychometrics, the application of item response model - a modern test theory - to psychometric problems, can help to overcome the concern about sample dependent estimates.

Figure 1: Item Calibration Using the Rasch Model (Item by Person Distribution Map)



Form A

```
Rasch Model Item Calibration Program -- RASCAL (tm) Version 3.51

ITEM BY PERSON DISTRIBUTION MAP
                                                      Numbers of
ITEMS                     PERSONS                   Items / People

        +  -4.0  +                                    0  /  0
        |  -3.8  |                                    0  /  0
        |  -3.6  |                                    0  /  0
      ##|  -3.4  |                                    3  /  0
        |  -3.2  |                                    0  /  0
        +  -3.0  +                                    0  /  0
        |  -2.8  |                                    0  /  0
 #######|  -2.6  |                                    8  /  0
        |  -2.4  |                                    0  /  0
     ###|  -2.2  |                                    4  /  0
    ####+  -2.0  +                                    5  /  0
        |  -1.8  |#                                   0  /  1
       #|  -1.6  |                                    1  /  0
      ##|  -1.4  |                                    2  /  0
  ######|  -1.2  |                                    7  /  0
     ###|  -1.0  |                                    4  /  0
   #####|  -0.8  |                                    6  /  0
     ###|  -0.6  |                                    4  /  0
   #####|  -0.4  |#                                   6  /  1
 #######|  -0.2  |#                                   8  /  1
 #######+<  0.0  +#                                   8  /  1
     ###|   0.2  |####                                4  /  3
 #######|   0.4  |##########                          8  /  7
      ##|   0.6  |######                              3  /  4
    ####|   0.8  |############                        5  /  9
   #####+   1.0  +##########                          6  /  7
     ###|   1.2 >|#######                             4  /  5
  ######|   1.4  |#######                             7  /  5
   #####|   1.6  |###########                         6  /  8
       #|   1.8  |########                            1  /  6
    ####+   2.0  +#######                             5  /  5
        |   2.2  |#                                   0  /  1
      ##|   2.4  |######                              3  /  4
      ##|   2.6  |###                                 2  /  2
        |   2.8  |                                    0  /  0
        +   3.0  +#                                   0  /  1
        |   3.2  |#                                   0  /  1
        |   3.4  |                                    0  /  0
      ##|   3.6  |                                    2  /  0
        |   3.8  |                                    0  /  0
       #+   4.0  +                                    1  /  0
+----+----+----+----+----------+----+----+----+----+
20   15   10   5               5    10   15   20
  Percent of                   Percent of
    Items                      Examinees

Summary Information:      Average      S.D.        Average     S.D.
                          Difficulty   Difficulty  ability     ability
(Theta Metric)            -0.00        1.60        1.22        0.83
(Scaled Score Metric)     100.0        14.5        111.1       7.6
```

Form B

```
Rasch Model Item Calibration Program -- RASCAL (tm) Version 3.51

ITEM BY PERSON DISTRIBUTION MAP
                                                      Numbers of
ITEMS                     PERSONS                   Items / People

         +  -4.0  +                                   0  /  0
         |  -3.8  |                                   0  /  0
         |  -3.6  |                                   0  /  0
       ##|  -3.4  |                                   3  /  0
         |  -3.2  |                                   0  /  0
         +  -3.0  +                                   0  /  0
  #######|  -2.8  |                                   8  /  0
         |  -2.6  |                                   0  /  0
        #|  -2.4  |                                   1  /  0
         |  -2.2  |                                   0  /  0
        #+  -2.0  +                                   1  /  0
       ##|  -1.8  |                                   2  /  0
       ##|  -1.6  |                                   2  /  0
      ###|  -1.4  |                                   4  /  0
     ####|  -1.2  |                                   5  /  0
        #|  -1.0  |                                   1  /  0
    #####|  -0.8  |                                   6  /  0
    #####|  -0.6  |#                                  6  /  1
   ######|  -0.4  |#                                  7  /  1
  #######|  -0.2  |#                                  9  /  1
 ########+<  0.0  +#########                         10  /  6
   ######|   0.2  |############                       7  /  8
      ###|   0.4  |#########                          4  /  6
    #####|   0.6  |######                             6  /  4
  #######|   0.8  |#######                            9  /  5
   ######+   1.0 >+#########                          7  /  6
       ##|   1.2  |#########                          2  /  6
      ###|   1.4  |##############                     4  /  10
       ##|   1.6  |######                             3  /  4
       ##|   1.8  |#########                          2  /  6
     ####+   2.0  +###                                5  /  2
     ####|   2.2  |###                                5  /  2
        #|   2.4  |#                                  1  /  1
         |   2.6  |                                   0  /  0
        #|   2.8  |                                   1  /  0
        #+   3.0  +                                   1  /  0
         |   3.2  |                                   0  /  0
         |   3.4  |                                   0  /  0
        #|   3.6  |                                   1  /  0
         |   3.8  |                                   0  /  0
         +   4.0  +                                   0  /  0
+----+----+----+----+----------+----+----+----+----+
20   15   10   5               5    10   15   20
  Percent of                   Percent of
    Items                      Examinees

Summary Information:      Average      S.D.        Average     S.D.
                          Difficulty   Difficulty  ability     ability
(Theta Metric)            0.00         1.48        0.94        0.76
(Scaled Score Metric)     100.0        13.4        108.5       6.4
```

Figure 1 shows the item by person distribution maps for each test form. The two distribution maps show similar profiles. By centering the scale on item difficulty, there was a mean average ability of

1.22 (sd=.83) for Form A and .94 (sd=.70) for Form B. These results show that the average ability of the examinees was higher than the mean item difficulty of either test forms.

The test and item analyses performed in this test piloting have important implications for NIE-CELT. One is the inclusion of quality test items to ensure validity of test score interpretation. In addition, the information obtained will be useful in determining cut-off scores for English Language proficiency expected of incoming NIE students.

Computerization of NIECELT: Working towards a prototype

Following the NIECELT team's research up to this point in time, a prototype of the NIECELT, which covers all the seven test domains, is currently being developed. It will be "computer-assisted", rather than completely "computer-based" in the sense outlined by Alderson (1988).[1] While at least five of the seven domains of NIECELT will be administered and scored entirely by computer, the *Reading Aloud & Speaking* (Domain 1) and the *Essay*, that is, spontaneous, continuous writing (Domain 7) are not wholly amenable to computer scoring. It has been widely recognized that "direct tests of oral and written production are, at present, simply not feasible in CBELT form." (ibid).

The NIECELT prototype will be subjected to further rigorous pilot testing on representative samples of student populations, and will be refined in accordance with the findings of the pilot tests. The prototype will incorporate digital video to introduce "the virtual invigilator" who will issue the test instructions and put the candidates at ease. There will also be provisions for candidates, once they log on to NIECELT, to be first familiar with the test tasks through practice items for all of the test domains they have been accordingly assigned to by the computer. Candidates will also have the benefit of being able to review their decisions in any part of the test taken.

Domain 1 will comprise two reading passages, one narrative and the other expository, which candidates will read aloud. Performance will be assessed on different criteria by trained examiners. Candidates will also be required to deliver a two-minute speech on an issue that arises from the oral reading of the expository text. The oral data will be captured on CDs or data files, which can be downloaded for manual scoring later.

Domain 2, which is listening-comprehension, will have candidates perform tasks that require them to follow some instructions and to listen to short texts - interviews, lectures, speeches, and conversations. The computer will assist by presenting the oral test instructions, providing appropriate visuals, title and questions to set the context and purpose for each listening task, and scoring the candidates' performance.

Domain 3, among other things, tests word order, sentence combining and paraphrasing, and recognition and correction of student-produced errors. The candidates' performance will be assessed by computer. Grammar, which is the mainstay of Domain 3, will also be tested more integratively in the cloze (in Domain 6).

Domain 4, is essentially a vocabulary test, and it assesses the candidates' ability to use appropriately close synonyms and opposites, as well as contextualized vocabulary through filling in the blanks of a passage with the aid of a list of words given at the head of the passage. This list will contain twice as many words needed for the blanks. The computer-enabled drag and drop feature for use with the filling-in-the-blanks section eliminates the possibility of spelling error in the task. Scoring is managed by the computer. The candidates' extent of lexical repertoire is also examined in other ways in the cloze in Domain 6.

Domain 5 has five medium-length texts culled from various genres for testing reading-comprehension. Each comprehension text and the attendant multiple-choice questions will be set in

[1] CBELT refers to "tests that are delivered by computer and also scored by computer" (Alderson, 1988:5).

a split-screen format on the computer to allow for easy question-to-text referencing. As before, scoring will be computer-assisted.

Domain 6 consists of a cloze test of about 50 blanks. Regular deletions are made at every fifth word, with the first letter of the word left intact in the blank. Exact word restoration is required for completing the cloze, and scoring is done by the computer.

Domain 7 is computer-assisted in so far as it dispenses written test instructions, questions, and graphics and statements as input for writing an expository essay and an argumentative essay respectively. The candidates' written performance will be assessed largely on their ability to write and organize material coherently rather than on their knowledge of the subject matter. The two essays, each to be between 350 and 500 words long, will be written by the candidates using the word-processing facility on the computer, and their written performance will be humanly marked on different assessment criteria.

The NIECELT team of researchers recognizes the full potential of computer technology as it exists on the market and also respects its limitations with regard to computerized testing. It will continue to upgrade NIECELT in the light of available technology but the cautionary fact must remain that while computers may afford "increased speed and possibly efficiency", they may have "nothing to offer testing in terms of increased validity." (Alderson, 1988:1)

References

Alderson, J. Charles (1988). *Innovation in Language Testing: Can the Micro-computer Help?* Lancaster: University of Lancaster

Alderson, J. Charles & Brian North (eds.) (1991). *Language Testing in the 1990s: The Communicative Legacy.* London: Modern English Publications.

Assessment Systems Corporation. (1994). *User's manual for the MicroCAT Testing System, version 3.5*, St. Paul, MN, Author.

Bachman, Lyle F. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, Lyle F. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests.* Oxford: Oxford University Press.

Brown, Annie & Tom Lumley (1991). *The University of Melbourne ESL Test: Final Report.* Melbourne: Language Testing Centre.

Milanovic, Micahel & Nick Saville (eds.) (1996). *Performance Testing, Cognition and Assessment: Selected Papers from the 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem.* Cambridge: Cambridge University Press.