

어문연구 제5권 1995년 2월

Toward Computer Adaptive Testing in English Proficiency Measurement

Jeong-ryeol, Kim*

Jeong-ryeol, kim. (1995). **Toward Computer Adaptive Testing in English Proficiency Meseasurement.** *Language & Literature Research*, 5, 97-121.

Computers became educationally more accessible to students and educators due to its ever decreasing price and ever increasing pieces of software. One of many areas computers are useful is computer adaptive testing. The potential benefits of computer- adaptive testing are reliability, repeatability, efficiency and standardization that are difficult to achieve with traditional paper- and-pencil tests. The testing was conducted on Korean junior high school students in collaboration with a teacher in the school using the adaptive testing software designed by the author. This paper will report on the design and implementation issues as to how one can start out to make this sort of software and how well the experimental results correspond to the results of the traditional paper-and-pencil tests.

1. Introduction

Computers became educationally more accessible to students and educators due to its ever decreasing price and ever increasing pieces of software. A whole new generation of students who grew up with computers and other electronic gimmicks are about to be in front of us, teachers. These students are at ease with flashing electronic games and multimedia software. Also, they are accustomed to the incessant

* 한국해양대학교, 교양과정부 조교수

stimuli transmitted through the computer screens or video games. These students are very likely to remain bored without a considerable level of stimuli and outside motivating factors. They will dislike and feel alienated by the traditional paper-and-pencil(P & P) media in the classroom. Thus, to attract students'attention, teachers must be prepared to incorporate some computational supplements into their regular teaching. One of many possible areas the author investigated is computer-assisted adaptive testing. The testing was conducted on Korean junior high school students using the adaptive testing software designed by the author.

Computer-adaptive testing(CAT) is used because people recognize its reliability, repeatability, efficiency and standardization that are difficult to achieve with traditional P & P tests. It also provides a highly individualized setting to take the test. Thus, CAT is very useful for the test mass-administered such as placement test, drivers' license exam and TOEFL. Also, as Tung(1986:27) points out, CAT based on continual estimates of each examinee's ability does not assign any questions inappropriately difficult to the examinees. This will produce desirable affective effects on the examinees, who will discover that the test items are always challenging but never much beyond their capabilities. The CAT program in this paper will start out at certain preset level of test question and accord to the responses given by students by lowering or raising the level of questions. The level of difficulty will move up and down until students respond a sufficient number of questions at which time the computer will set the level of the following the students their language proficiency. The software is originally designed to test

the reading proficiency but in this experiment it is expanded to cover grammar and vocabulary. It also has potential to incorporate listening test by using some sort of sound card. All the records of how students answered will be kept for teachers to locate the problem area of the students. If the examiner specifies the purpose of certain question and makes an index along with the main test questions, (s)he will find that the test result can serve a good diagnostic tool as well.

The test result shows that students who performed well on the traditional testing does well on the computer testing too. No surprise! However, it took less time and provided more individualized environment. Also, since the questions are shown by the degrees of difficulty, it lessens the chances of students' guessing on the questions especially when they are confronted with questions far more difficult than those they can handle.

There are problems which teachers interested in this area must be aware of. One is the anxiety factor when a student is not familiar with computer. Another is scaling the questions by their degree of difficulty. Some of the solutions will be discussed during the course of the main discussion.

This research needs to incorporate the latent trait theory also called item response theory (Hambleton & Swaminathan 1985, Lord & Novick 1968, Rasch 1980 and Wright & Stone 1979) which includes estimating performance of persons and items on a single difficulty continuum (one parameter model), a discrimination continuum in addition to the difficulty continuum, and a measurement error continuum such as guessing. These differing numbers of parameters of item and person performance will help

measure up the CAT results.

2. Design Issues

The foundation of CAT is how well test questions approximate to the ability of examinees. Wainer's track-and-field metaphor (1983) illustrates this concept. For persons whose high-jumping range is approximately 1.5 meter, hurdles less than 1 meter or higher than 2 meters would contribute very little to the precise measurement of high jumping ability. But precision would be added by providing hurdles close to 1.5 meter height. Providing this precision to examinees must be always the prime concern of CAT test designer.

To make a successful CAT the author must consider three different dimensions: question data base, mode of presentation and affective effects on the examinees. These three dimensions contribute from different angles to make the test successful.

2.1 Question Data Base

Test questions must be pretested to the same target of population to measure the appropriateness of test questions and also to create a scalar system in terms of their relative difficulty among the employed test questions. Test questions must be stored in a format for examiners to increment additional questions easily and for examinees to retrieve any appropriate level of questions to their proficiency efficiently. To restate this general goal, Henning(1991:210-212) suggests that the data base must include the following five characteristics:

- (1) Items included in an item bank have been preadministered and analyzed for use with the same intended population of examinees.
- (2) The items in a CAT item bank are stored together according to some system that permits ready retrieval and rapid presentation for testing purposes.
- (3) Items in an item bank have been calibrated or placed on some measurement continuum according to a uniform scalar metric.
- (4) The possibility exists of increasing or decreasing the number of items in the item bank without destroying the usefulness of the prior remaining items or changing the meanings of the item classifiers.
- (5) Responses to items in a CAT item bank that are used to generate a generalized measure of ability must cumulatively reflect that ability.

In the experiment, total 400 test questions are collected from ten previous city wide exams and those questions were pretested 10 different times to different group of 190 students in four different classes and the test results were analyzed to determine their relative difficulty by the rate of students who correctly answered the question. The test questions are classified into three different levels based on the rate of students who correctly answered the question. When the passing rate is below 30%, the question is labeled as difficult, when the passing rate is between 30% and 70%, the question is classified as average and finally when the passing rate is greater than 70%, the question is labeled as easy. Any questions which meet less than 5% of passing rate and more than 95% of passing rate are eliminated

to secure the distinctive power of each question. After excluding questions not meeting the required passing rate, the test includes 350 questions of which 60 are difficult, 120 are average and 170 are easy. The areas of testing include reading comprehension, grammar and vocabulary. The testing focus was given with a special reference to the 6th curriculum reform which emphasizes the functional and notional skill getting through the secondary school English classes in Korea.

Since the data are stored in such a format as one question on one card in Macintosh HyperCard, it is extremely easy to increment any additional questions or delete any question later found to be inappropriate. Examiner can create a card which already contain all the necessary field and buttons as background and fill the blank text field with additional questions. When an examiner create a card, (s)he will be given a card as in fig 1 which consists of two text fields and 4 radio buttons. The upper text field is for the text and the bottom text field is for the question and multiple choices, and the radio buttons at the bottom are four choices for examinees to make.

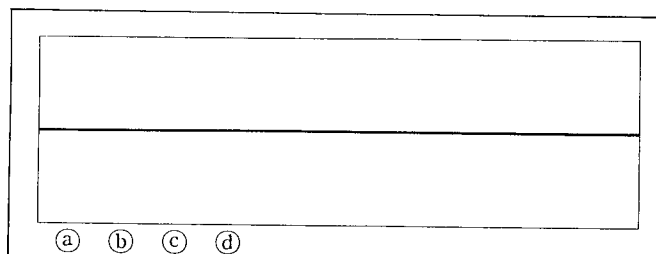


Fig. 1 Background of a new card for additional test question

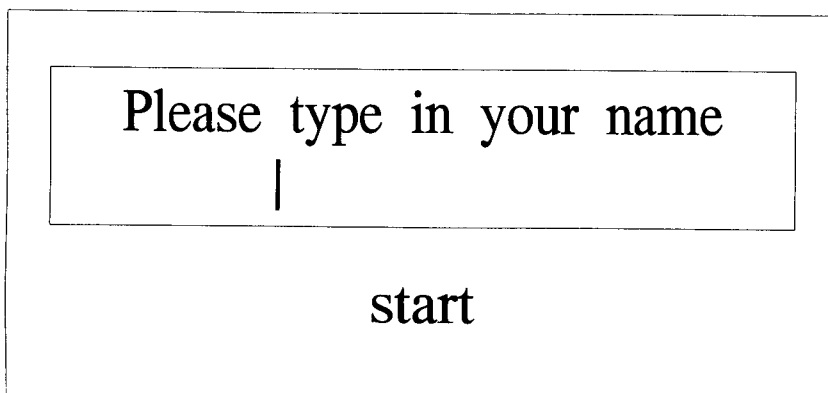
Also, if necessary, examiners can delete the card which contain

ill-fit question for certain test purpose without any difficulty.

2.2. Mode of Presentation

One of major concerns in computer-assisted language testing (CALT) in general is the validity of this test compared with the traditional P & P test. Most of the current functional CALTs rely on analysis results derived from this traditional mode of testing, even though the targeted application of these items is via the medium of the computer (Henning 1986; Hicks 1986; Kaya-Carton, Carton & Dandonoli 1991; Larson 1987). The debate on the test validity centers around the possibility that test items may function differently depending on different modes of presentation. Green(1988) has provided evidence, however, that item difficulty estimates for both verbal and mathematical items are invariant across computer and P & P presentation modes. Still, the validity issue does not fade away since Green's test does not cover every possible test format and the possibility exists that certain item formats or examination tasks in CALT mode may result in different performances among students compared to the traditional mode.

As testing begins, the student is asked to sign up by typing in his/her name at the second line in the field where the cursor is blinking and press the start button using mouse to start the test.

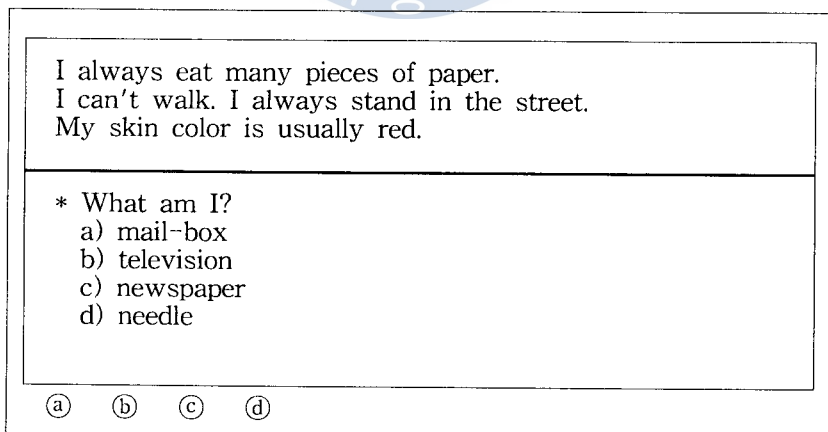


Please type in your name
|

start

Fig. 2 Initial Card for the CAT

Testing begins with the presentation of an easy question from the question data base, and scale up when students answer correctly and the responses meet the acceptable error rate of around 0.01. Student will be given a test question appeared in fig 3 as follows:



I always eat many pieces of paper.
I can't walk. I always stand in the street.
My skin color is usually red.

* What am I?
a) mail-box
b) television
c) newspaper
d) needle

(a) (b) (c) (d)

Fig. 3 Example of an easy question

The question may be presented in Korean if it actually is too difficult for student to understand the direction. Student selects one of the given four buttons(a, b, c and d) and his/her choice will be recorded and reflected for a next possible step to take.

In selecting test items in CAT, Stevenson and Gross(1991) applies the Bayesian strategy which brings up easier questions when the student answered incorrectly and harder items when the student answered correctly.¹ The strategy applied in the testing procedure for this paper is very similar to the Bayesian strategy which the Standard Error of Measurement(SEM) starts at 1 and is gradually reduced to 0.01 as the testing progresses. When the SEM hits the acceptable error rate, our test does either introduce an easier level of test item or a more difficult set of test items or determine the student's proficiency level depending on how the student answered to the previous series of questions presented.

Choi E S's test result on 1993.10.18	3,a.,a.,r,High 39
1,a.,a.,r,Novice 63	3,d.,d.,r,High 10
1,b.,d.,w,Novice 107	3,c.,c.,r,High 11
1,c.,c.,r,Novice 56	3,a.,a.,r,High 9
1,d.,d.,r,Novice 28	3,a.,a.,r,High 28
1,c.,c.,r,Novice 16	3,b.,b.,r,High 36
1,c.,c.,r,Novice 100	3,a.,b.,w,High 14
2,d.,d.,r,Intermediate 42	3,c.,b.,w,High 49
2,c.,c.,r,Intermediate 64	2,c.,c.,r,Intermediate 55
2,c.,c.,r,Intermediate 96	2,c.,c.,r,Intermediate 98
2,b.,b.,r,Intermediate 20	2,c.,d.,w,Intermediate 27
2,c.,d.,w,Intermediate 16	2,b.,c.,w,Intermediate 47
2,c.,c.,r,Intermediate 88	2,b.,a.,w,Intermediate 28
2,b.,c.,w,Intermediate 115	2,a.,d.,w,Intermediate 70
2,d.,d.,r,Intermediate 4	2,b.,b.,r,Intermediate 103
2,d.,d.,r,Intermediate 8	2,c.,c.,r,Intermediate 108
3,a.,d.,w,High 48	Advanced Marker
3,b.,a.,w,high 50	Time Used: 5650.125
3,b.,c.,w,High 26	

Fig. 4 Student record

Student responses were recorded and stored on a hard-disk student record that includes student name, test date, item identification number and its difficulty category, keyed (correct) response, the student's response, duration of the test in ticks, number and proportion of items correct, and the likelihood of student's level as shown in fig 4.

2.3 Affective effects

Cambre and Cook(1985), and Campbell(1986) have focused on defining and developing measures of computer anxiety. Researchers believe that prior computer exposure and familiarization may predictably serve to reduce such anxiety(Gressard & Loyd 1984; Raub 1981). There is some further evidence that providing immediate feedback after each CAT item encounter may actually serve to heighten anxiety beyond what would be present if feedback were denied until the end of the test (Wise, Plake, Eastman, Boettcher & Lukin 1986). On the other hand, it is argued that the usual CALT process of presenting items that are matched to the ability of the examinee would serve to reduce anxiety below the levels experienced when a conventional test presents some items that are too difficult for the examinee (Henning 1991). Madsen(1991) suggests that the CAT promises to instill confidence that examinee's skills are being evaluated with precision and objectivity. At present there is no sufficient evidence that computer anxiety exists to the level of invalidating the CALT results. Also, by adding warm-up questions before the main test, the anxiety effect can be considerably diminished due to the familiarity with the test mode.

Another affective effect with CAT is that it is less threatening mode of testing. There will be no proctor staring at students to check whether or not they cheat. Since all the students do not progress at the same rate of speed, it is also possible for teachers to build in a flexible testing schedule, allowing students to take their exams when they feel ready to do so.

3. Implementation Issues

HyperCard running on Macintosh is employed to implement the characteristics described in the previous sections. HyperCard is chosen due to its relative ease in scripting and all the necessary parts and pieces in place which meet the experimental needs.

3.1 Flow Chart

As shown in the following flow chart, an examinee signs up his/her name and press the start button. The computer will then display one randomly chosen question from the easy question pool to start. The examinee responds to the question by pressing one of the four buttons, each representing one of the four multiple choices. The response is immediately evaluated against the answer key and counted for or against the examinee's proficiency record. The number of questions at certain level shown to the examinee is counted and serves as a determining factor along with the number of correct responses whether or not to show the next level of questions. The student ability level counter is to check whether or not enough tries have been attempted at certain level to set the level as the student's proficiency level. The test ends if it satisfies one of the following

two conditions: (1) There is no more level to adjust to. That is, the examinee's level is either advanced or novice. (2) The examinee's ability does not go over or under the current level. Enough tries have been attempted to establish the proficiency level of the examinee's. Also, when the examinee presses the start button, the timer starts to count the time and finishes as soon as the proficiency level is determined.

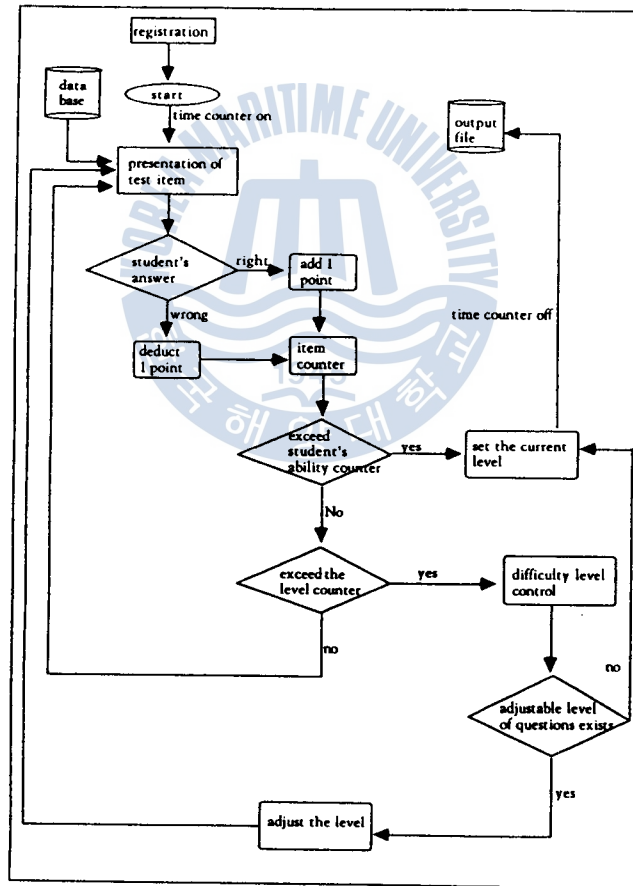


Fig. 5 Flow Chart

3.2 Scripting

In this section, the scripting is reviewed in the order of presentation given in the flow chart. First, let's look at what happens when an examinee presses the start button.

Script 1. Start

```
on mouseUp
  global Showed, name, InterCounter, NowPos, LocalCount
  if line 2 of cd fld "name" is empty then
    answer "Type in your name, please."
  else
    put line 2 of cd fld "name" into name
    put 1 into InterCounter
    put random (number of lines of cd fld "lv 1-1" of cd
"setting")
into StartNo
go cd line StartNo of cd fld "lv 1-1" of cd
"setting"delete line StartNo of cd fld "lv 1-1" of cd
"setting"put 1 into NowPos put 1 into LocalCount
end if
end mouseUp
```

It checks if the examinee signed up his/her name in the given name field and sends a message "type in your name, please." if the examinee did not already sign up. Otherwise, it puts the examinee's name into the global variable "name" and starts counter from 1 and randomly chooses one card from the easy questions pool and go to the card and sets the card unavailable in order not to be shown to the same examinee again.

When the randomly chosen card is displayed, the examinee sees the text and a question about the text. The examinee is expected to respond to the question by pressing one of the four multiple choices. Pressing one of the four buttons executes the Script 2.

Script 2. Student Response

```

on mouseUp
  myAnswer the short name of me, short name of this
  cdend mouseUp

on myAnswer Choice, CdName
  global RecordLn, InterCounter, Direction, adjustglobal
  PrevPos, NowPos, LocalCount, FinishNum

  put fld "answer key" & "." into GoodAnswer
  if NowPos > 0 and NowPos < 4 then
    if Choice GoodAnswer then
      put Direction -1 into Direction
      put NowPos & "," & GoodAnswer & "," & Choice & "," &
      "w" & "," & CdName into line InterCounter of RecordLnelse
      put Direction + 1 into Direction
      put NowPos & "," & GoodAnswer & "," & Choice & "," &
      "r" & "," & CdName into line InterCounter of RecordLnend
    if
      if abs (Direction) >= adjust then
        if Direction < 0 then put NowPos - 1 into NowPoselse put
        NowPos + 1 into NowPos
        put 0 into Direction
      end if
  
```

```
-- Finish testing if the local frequency exceeds the preset
number
if NowPos = PrevPos then put LocalCount + 1 into
LocalCountelse put 0 into LocalCount
put NowPos into PrevPos
if LocalCount > FinishNum then
Record (RecordLn)
exit myAnswer
end if
end if
-- how often have you visited the level?
if NowPos < 4 and NowPos > 0 then
put LocalCount into item 2 of line NowPos of cd fld
"fluency" of cd "Record"
end if

if NowPos = 0 then
Record (RecordLn)
answer "Could you work harder?"
else
if NowPos > 3 then
-- 3 is the number of lines of cd field "fluency label"-- of
cd "setting"
Record (RecordLn)
answer "Relax until we have more difficult materials!!"else
put InterCounter + 1 into InterCounter
cue (NowPos)
end if
end if
```

```
end myAnswer
```

When student chooses one of the four given multiple choices, the answer and card name are passed on to the action called "myAnswer". The action "myAnswer" checks if the answer is correct by comparing the answer key and the result is counted toward or against the counter called "direction". There are two directions + and - which indicates whether to introduce an upper level question or a lower level question. When the counter exceeds more than 3 which approximates to the SEM 0.01 in its error rate, then the CAT introduces an upper level question or a lower level question depending on the direction. However, if there is no more adjustable text, then the CAT sets the current level as the student ability level and the script 4 runs and sets student level. Otherwise, the following script takes over to present the next question.

Script 3. Presenting a question

```
on cue RefNo
  global RecordLn
  -- dehilite all the choice buttons
  set hilite of bg btn "a." to false
  set hilite of bg btn "b." to false
  set hilite of bg btn "c." to false
  set hilite of bg btn "d." to false

  if RefNo = 1 then
    put random (number of lines of cd fld "lv 1-1" of cd
"setting") into StartNo
    go cd line StartNo of cd fld "lv 1-1" of cd "setting"delete
```



```

line StartNo of cd fld "lv 1-1" of cd "setting"else if RefNo
= 2 then
  put random (number of lines of cd fld "lv 2-1" of cd
"setting") into StartNo
  go cd line StartNo of cd fld "lv 2-1" of cd "setting"delete
line StartNo of cd fld "lv 2-1" of cd "setting"else if RefNo
= 3 then
  put random (number of lines of cd fld "lv 3-1" of cd
"setting") into StartNo
  go cd line StartNo of cd fld "lv 3-1" of cd "setting"delete
line StartNo of cd fld "lv 3-1" of cd "setting"end if
end cue

```

Script 3 first cleans up the marked button in the test bed so that an examinee is not influenced by the marked button. Then, it checks whether to scale up or down the question level. It may stay at the same level. In any case, it picks up a randomly chosen card from the requested pool of difficulty level and present it to the examinee.

As mentioned earlier, as the examinee's ability level is determined, the script 4 runs to keep a record of his/her testing.

Script 4 Setting the proficiency level

```

on Record text
  global name
  put cd fld "fluency" of cd "record" into MyRecordrepeat
with i=1 to number of lines of MyRecord
  put item 2 of line i of MyRecord & "," after
  Compared

```

```

put max(Compared) into MostVisit
end repeat
repeat with j=number of lines of MyRecord down to lif
MostVisit is in line j of MyRecord then
put j into CurrentLv
exit repeat
end if
end repeat
put Stat (CurrentLv, text) into yourLevel
answer "Would I offend you if I say your level is" &&
yourLevel & "?"
put name & "'s test result on " & the date & return into
cd fld "Record fld" of cd "Record"
put text & return after cd fld "Record fld" of cd
"Record"put yourLevel after cd fld "Record fld" of cd
"Record"Answer "Do you want to see your record in
detail?" with "no" or "yes"if it = "yes" then go cd "record"
end Record

```

```

function Stat Index, Base
put 0 into RightCount
repeat with i=1 to number of lines of Base
if item 1 of line i of Base = Index then
put item 4 of line i of Base & return after Poolif item 4 of
line i of Base = "r" then
put RightCount + 1 into RightCount
end if
end if
end repeat
if RightCount / number of lines of Pool * 100 >= 80

```

```

thenput "high" into Marker
  else if RightCount / number of lines of Pool * 100 < 80
and RightCount / i * 100 >= 50 then
  put "med" into Marker
  else if RightCount / number of lines of Pool * 100 < 50
then put "low" into Markerput item 1 of line Index of cd fld
"fluency" of cd "Record" into Levelreturn (Level &&
Marker)
end Stat

```

Script 4 sets the proficiency level of the examinee's and further calculates how well (s)he did on the test in such a way that if the number of correct answers from the set proficiency level equals to or exceeds 80%, then the examinee is marked as "high", and if it is between 50% and 80%, "mid" and finally if it is less than 50%, "low". Also, it records the result of the examinee's test.

4. Experimental Results and Conclusion

This test was given to 35 students from Puil Junior High School in Pusan, South Korea. The table 1, experimental results, consists of the sequential number starting from the top student in the first column, number of questions in CAT presented to examinees in the second column, the level each student achieved from CAT testing in the third column, number of questions presented in the P & P testing and finally the scores earned by students in the last column. The scores in the last column is from the previous P & P English test basically covering the same scope of test range. The test result demonstrates that (1) a

smaller number of questions were presented and used to check the student's proficiency level and (2) there is a high correlation rate between CAT test results and the student's score earned from the traditional P & P testing, though there are some deviating element such as #8, #19 and #31. At this point, it is not ready to interpret as to why and how this deviation exists, however, this does not invalidate the correlation rate existing between two different modes of testing as shown in Graph 1 of the correlation graph.

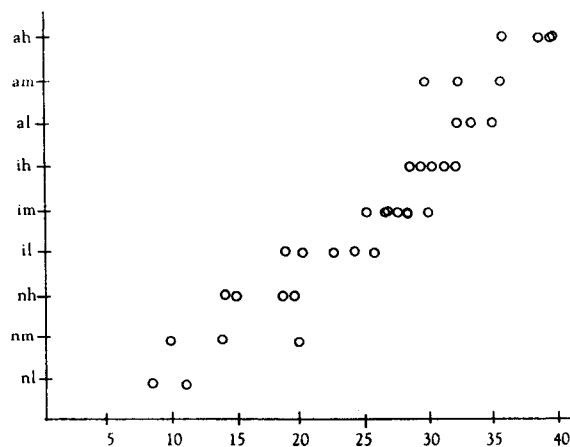
Table 1. Experimental Results

#	# of Questions (CAT)	Level	#of Questions (P&P)	Score
1	9	Advanced High	40	40
2	9	Advanced High	40	38
3	11	Advanced High	40	40
4	11	Advanced Mid	40	36
5	14	Advanced Mid	40	36
6	16	Advanced Mid	40	32
7	16	Advanced Low	40	30
8	18	Advanced Low	40	35
9	34	Advanced Low	40	33
10	10	Intermediate High	40	32
11	17	Intermediate High	40	30
12	11	Intermediate High	40	31
13	13	Intermediate High	40	28
14	30	Intermediate High	40	29
15	15	Intermediate Mid	40	28
16	18	Intermediate Mid	40	26
17	26	Intermediate Mid	40	26
18	14	Intermediate Mid	40	27
19	16	Intermediate Mid	40	30
20	15	Intermediate Mid	40	25
21	23	Intermediate Mid	40	28
22	14	Intermediate Low	40	26

23	16	Intermediate Low	40	22
24	15	Intermediate Low	40	24
25	14	Intermediate Low	40	18
26	17	Intermediate Low	40	20
27	17	Intermediate Low	40	15
28	10	Novice High	40	18
29	11	Novice High	40	14
30	20	Novice High	40	19
31	28	Novice Mid	40	20
32	19	Novice Mid	40	10
33	10	Novice Mid	40	14
34	7	Novice Low	40	12
35	8	Novice Low	40	8

As mentioned above, the correlation rate between the CAT and the P & P test is very high, though there are few outlying scores which do not reflect this. This indicates the mode of presentation in the test, at least CAT and P & P, does not make much difference in the student performance.

Graph 1. the correlation graph.



In conclusion, this experiment provides a method as to how to design and implement a CAT for English as a foreign language. The test result from this experiment is encouraging so that this mode of test can be validated to test the students' level of English proficiency. If CAT is selected as a mode of language testing, the following advantages will follow:

(1) It provides a highly individualized testing environment in which student takes the test whenever (s)he feels ready. Everybody acquires a foreign language skill in a different rate of speed and deserves some respect in his/her phase of acquiring the target language skill. The traditional P & P method given simultaneously at the same time disregards individual student's difference to accomplish a uniform measurement.

(2) The test administration can be simplified a great deal since it does not require anybody to proctor the exam. The only limit is number of monitors available to meet the number of examinees. The hours of testing can be stretched around the clock if it is necessary.

(3) The feedback is immediate in such a way that when an examinee finishes the test, (s)he already knows which level of proficiency (s)he is at.

(4) It is especially appropriate mode of testing when the test needs mass-administering. For example, language placement testing or driver's license exams can be a good example to employ this mode of testing. All the examinees can get the immediate result printed on certain paper to hand in for the

further process.

Notes

1. A Bayesian item selection procedure selects items on the basis of the precision (or standard error of measurement) of an examinee's estimated ability. Testing begins with the Standard Error of Measurement (SEM) set at 1.0. By administering items of varied difficulty, an examinee's ability estimate is refined and the SEM shrinks. Testing ends when the SEM becomes sufficiently small (usually around 0.01).

References

- Cambre, M. A., & D. L. Cook (1985). Computer anxiety: Definition, measurement, and correlates. *Journal of Educational Computing Research*, 1(pp. 37-54).
- Campbell, N. J. (1986). *Technical characteristics of an instrument to measure computer anxiety of upper elementary and secondary school students*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.
- Chung, Dong-soo (1995). On Using multi-media for English language education. *Ungyong Enehak (Applied Linguistics)*, 8(pp. 181-206).
- Goodman, D. (1990) *HyperCard 2.0*. New York: Bantam Books.
- Green, B. (1988) Construct validity of computer-based tests. In H. Wainer and H. I. Braun (Eds.), *Test validity* (pp. 77-86). Hillsdale, NJ: Erlbaum.
- Gressard, C. & B. H. Loyd (1984). *An investigation of the effects of math anxiety and sex on computer attitudes*. Paper

- presented at the annual meeting of the American Educational Research Association, New Orleans.
- Hambleton, R. K. & H. Swaminathan(1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.
- Henning, G.(1986). Item banking via DBASE II: The UCLA ESL proficiency examination experience. In C.W. Stansfield (Ed.), *Technology and language testing*(pp. 69-77). Washington, DC: TESOL.
- Henning, G.(1991). Validating an item bank in a computer-assisted or computer-adaptive test: using item response theory for the process of validating CATS. In P. Dunkel(Ed.), *Computer-assisted language learning and testing*(pp209-221). New York: Harper Collins Publisher.
- Hicks , M.(1986). *The TOEFL computerized placement test: Adaptive conventional measurement*. TOEFL Research Report # 31. Princeton, NJ: Educational Testing Service.
- Kaya-Carton, E., Carton, A. & P. Dandonoli(1991). Developing a computer-adaptive test of French reading proficiency. In P. Dunkel(Ed.), *Computer-assisted language learning and testing* (pp. 259-284). New York: Harper Collins Publisher.
- Larson, J. W.(1987). Computerized adaptive language testing: A Spanish placement exam. In K. M. Bailey, T. L. Dale and R. T. Clifford(Eds.), *Language testing research*(pp. 1-10). Monterey, CA: Defense Language Institute.
- Lim, Chang-keun(1994). Adaptive language testing using computer. *Journal of language sciences*, vol. 1(pp. 225-246). Pusan: Southeast Linguistic Society of Korea.
- Lord, F. M., & M. R. Novick(1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

- Madsen(1991). Computer-adaptive testing of listening and reading comprehension: The Brigham Young University approach. In P. Dunkel(Ed.), *Computer-assisted language learning and testing*(pp237-258). New York: Harper Collins Publisher.
- Rasch, G.(1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.
- Raub, A. C.(1981). *Correlates of computer anxiety in college students*. Unpublished doctoral dissertation, University of Pennsylvania.
- Tung, P.(1986). Computerized adaptive testing: Implications for language test developers. In C. W. Stansfield(Ed.), *Technology and language testing*(pp. 11-28). Washington, DC: TESOL.
- Wainer, H.(1983). On item response theory and computerized adaptive tests. *Journal of Educational Measurement*, 14(pp 181-196).
- Wise, S. L., Plake, B. S., Eastman, L. A., Boettcher, L. L. & M. E. Lukin(1986). *The effects on test performance and anxiety of using examinee-selected item ordering in a computer-administered mathematics test*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.
- Wright, B. D., & M. H. Stone(1979). *Best test design: Rasch measurement*. Chicago: Mesa Press.

