# 行政院國家科學委員會專題研究計畫成果報告
## 中學生資訊科技之網路學習與評量系統之研究　子計劃二：網路化創造性學習環境之可行性研究(2/3)
## DIYexamer : A Web-based Multi-Server Testing System　with Dynamic Test Item Acquisition and　Discriminability Assessment

計劃編號：NSC 89-2520-S-009-006
執行期限：88 年 8 月 1 日至 89 年 7 年 31 日

主持人：林盈達 教授　　交通大學資訊科學系

**Abstract**

This paper presents a novel network CAT system, DIYexamer (Do-It-Yourself Examer). It has three features that differentiate it from existing CAT systems: student DIY items, item-bank sharing, and automatic assessment of item discriminability. DIYexamer accepts test items contributed form teachers as well as students, and allows limited item sharing between item-banks possibly maintained by different organizations. An algorithm is applied dynamically to assess the discriminability of items in item-banks in order to filter out less qualified contributions, hereby assuring the quality of stored items while scaling up the size of item-banks.

**Keywords** : Computer Assisted Testing, Test Evaluation,Test Acquisition, Discriminability, Distant Learning

### 1 Intorduction

Computer-assisted Testing (CAT) or Computer-based Testing (CBT), the use of computers for testing purposes, has a history spanning more than twenty years. The documented advantages of computer administered testing include reductions of testing time, an increase in test security, provision of instant scoring, and an individualized adaptive testing environment [1][2][3][4]. Three categories of CAT are currently employed: standalone packages, test centers and networked systems. Regardless of which CAT system is employed, a critical issue in developing CAT is the construction of a test item-bank. Traditionally, asking teachers and content experts to submit items generates the item-bank. Three major drawbacks of the traditional method can be observed:

1) Limitation of item amount: Teachers and content experts tend to have similar views on the test subject. That is, in a given field vital subject matter might be confined. Therefore, although more teachers and content experts are invited to contribute test items, the total number of distinct items remains low.
2) Passive learning attitude: Students are conventionally excluded from the creation of tests. In a typical computer-assisted testing system, teachers generate tests, the system presents test sheets and students then complete the tests. That is, within the system of testing, they play a passive role, and are not afforded the opportunity to conduct “meta-learning” or “meta-analysis.”
3) No guarantee on item quality: Permitting students to generate tests may be a possible solution to the aforementioned problems. However, this raises a new problem: quality assurance and ensuring that the tests are worth storing and used for further tests. Even when the whole item-bank is contributed by teachers and content experts, ways to dynamically assess and filter test items are needed.

### 2 The Diyexamer Solution

The DIYexamer[5] provides a web interface for users to remotely control and operate the system. Three types of users are supported: administrators, teachers, and students. It allows students to contribute test items, and provides an effective means of verifying the discriminability of these items. Three main ideas are as below:

1) Item DIY by students: DIYexamer allows students to generate test items into the item-banks online as Fig 1. Teachers can query these items generated by students as Fig 2. In addition to rapidly increasing the total number of items in an item-bank, this feature also encourages students to develop *meta-learning*, i.e. *creative learning*. In order to submit tests, students must thoroughly study the learning materials, develop higher-level overviews of the materials, and practice cognitive and creative thinking.

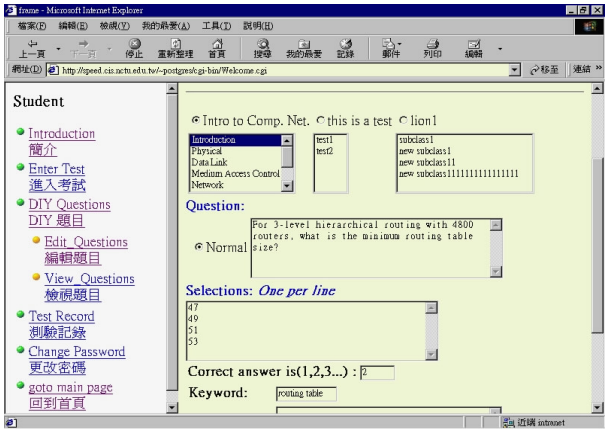


Fig 1: Students generate items into the item-bank

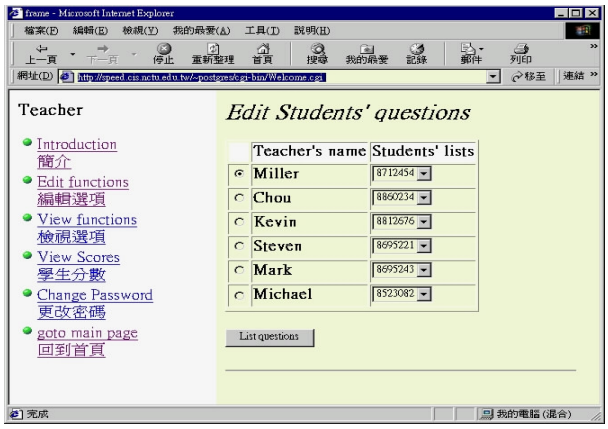


Fig 2: Student DIY items as queried by teachers

2) Assessment of item discriminability: DIYexamer provides an item-discriminability assessment method to ensure the quality of the stored items. In addition to ensuring the internal consistency of existing test items, this method also continuously and dynamically screens additional new items in the item-bank. Fig 3 shows the average item discriminabilities of several item-banks.

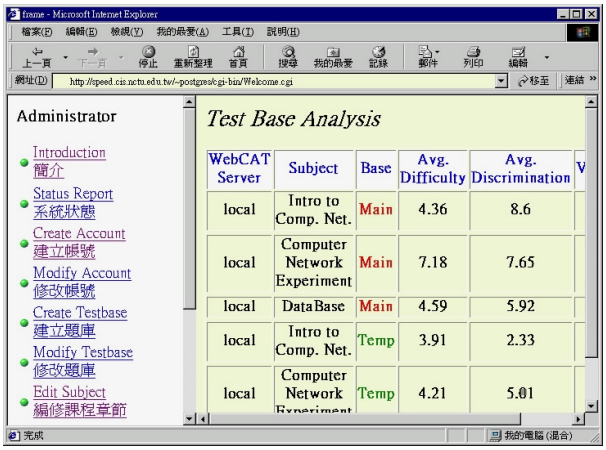


Fig 3: Average item discriminabilities of item-banks

3) Item-bank sharing: DIYexamer, a scalable multi-server system, connects many item-banks stored in different servers. Therefore, via the Internet, more items can be accessed and shared. The sharing is limited and controlled in a sense that a server issues a request, describing the criteria of a test item it requests, to another server. A server does not open up its item-bank for unlimited access.

Additional advantages have been identified and include the facts that since DIYexamer provides a real-time on-demand generation of test-sheet function, cheating is avoided. Also, DIYexamer provides an item cross-analysis function to which the degree of difficulty for each test as well as the entire test base can be accurately measured.

## 3 Discriminability Assessment Of Diyexamer

When selecting sample students, only those whose scores have large gap with the average score should be considered. Accordingly, those with the top 30%, in terms of range, scores are defined as “high-score group (H’)”, while those with the bottom 30% scores are defined as “low-score group (L’)”.

To show the different criteria and effects of choosing samples in the traditional method and DIYexamer method, Fig.4 depicts the score distribution in a test. In this example, the highest score is 92, the lowest score is 34, and the average score is 69. The “high rank score group” and the “low rank score group” are chosen according to these two methods. Take student X as an example, the score of X is 66, which differs only 3 points from the average score. The associated information of X should have little, if not none, referential value in computing item discriminability. However, X is chosen as a sample in the high rank group in the traditional method. This fallacy results from using rank group, in terms of count, as the criterion of choosing samples. In DIYexamer, X is not chosen since score group, in terms of range, rather than rank group is used. Only those with large gap with the average score are chosen as samples.

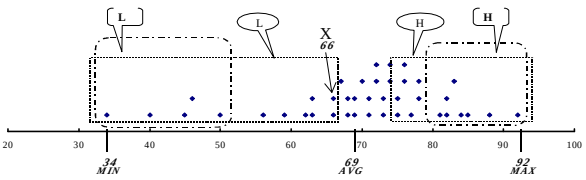


Fig 4: Comparison of samples taken in the traditional method and DIYexamer method

For different samples to have different impacts on discriminability, a referential value with

respect to an item is generated for each student selected as a sample. We first define the item discriminability as the average of all associated referential values, as shown below:

$$\text{Discriminability} = \frac{\text{Sum of the referential values of sampled students}}{\text{Number of sampled students}}$$

Since the referential values depend on students' scores, the referential values are computed according to the ratio of correct and incorrect answers of the sampled students. The ratios of correct and incorrect answers are defined as follows:

$$\text{Ratio of correct answer} = \frac{\text{Number of items answered correctly}}{\text{Number of items on the test}}$$

$$\text{Ratio of incorrect answer} = \frac{\text{Number of items answered incorrectly}}{\text{Number of items on the test}}$$

According to Table 1, the referential value of a student correctly answered an item is the ratio of correct answer of the student. Alternately, the referential value of a student incorrectly answered an item is the ratio of incorrect answer of the student. This policy comes from the fact that an item should have increased discriminability if correctly answered by a competent student, while rendering decreased discriminability if correctly answered by a less competent student. In this way, a competent student contributes large referential value to a correctly answered item and small referential value to an incorrectly answered item, and vice versa.

## 4 Evaluation Of The Discriminability Assessment In Diyexamer

The fairness and performance of DIYexamer was evaluated. We conducted an experiment where 10 students took the test on-line using DIYexamer with 10 items. Table 2 summarizes the test results. Fig 5 shows the score distribution of the experiment. Discriminability for each item is computed using both the traditional method and the DIYexamer method. However, the discriminability originally falls between -1 to 1 using the traditional method, while falling between 0 to 1 using the DIYexamer method. To compare these two methods, both two ranges of discriminability are then normalized to 0 to 10, as shown in Fig 6.

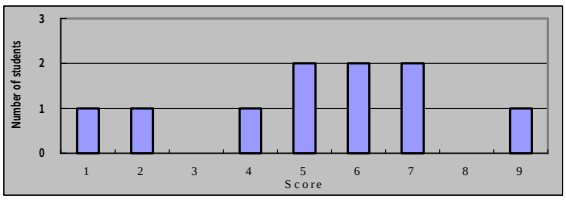


Fig 5: Score distribution of the test experiment

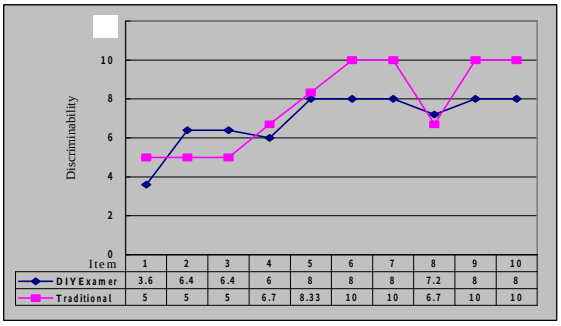


Fig 6: Comparison of item discriminability

TABLE 1: Principle to compute the referential value of a student with respect to an item

| Student | Answer | Item discriminability | Referential value to compute discriminability |
|---|---|---|---|
| Competent (With high ratio of correct answer) | Correct | High | Ratio of correct answer |
| | Incorrect | Low | Ratio of incorrect answer |
| Less competent (With low ratio of correct answer) | Correct | Low | Ratio of correct answer |
| | Incorrect | High | Ratio of incorrect answer |

TABLE 2: Result of the test experiment

| | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 | Item 7 | Item 8 | Item 9 | Item 10 | Number of correct answers(score) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| student1 | 1 (correct) | 0 (incorrect) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| student2 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| student3 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 4 |
| student4 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 5 |
| student5 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 5 |
| student6 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 6 |
| student7 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 6 |
| student8 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 7 |
| student9 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 7 |
| student10 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 9 |

## 5 Conclusion

This paper has presented a novel architecture for a networked CAT system, DIYexamer. It supports item DIY by students, item-bank sharing, and item discriminability assessment.

For discriminability assessment, new calculation formula were proposed. When compared with the traditional assessment scheme, the main difference is that the top and the bottom 30% of the *score* group, in terms of *range of scores* were selected rather than the *rank* group, in terms of *count of students*. Thus, item discriminability is more accurately reflected particularly when the tested students have close scores.

Item-bank sharing and item DIY by students has increased both the *amount* and the *variety* of questions in item-banks. Item DIY by students promotes *creative learning* within students, while automatic discriminability assessment assures better quality than traditional CAT systems.

A questionnaire was used to survey subjective attitudes of students about DIYexamer. As shown in Table 3, the outcome revealed that most students were interested in item DIY.

The technique proposed herein is useful in general tuition not only to improve the quality of test items and fairness; but also to save time from generating questions and computing scores. We recommend that DIYexamer be popularized to schools.

## 6 References


[1] C. V. Bunderson, D. K. Inouye, and J. B. Olsen, "The four generations of computerized educational measurement," in Educational measurement (3rd ed.), R. L. Linn, Ed. New York: American Council on Education—Macmillan, pp 367-407, (1989).

[2] S. L.Wise and B. S. Plake, "Research on the effects of administering tests via computers," Educational Measurement: Issues and Practice, vol. 8, no. 3, pp. 5-10, (1989).

[3] A. C. Bugbee, Examination on Demand: Findings in Ten Years of Testing by Computers 1982-1991. Edina, MN: TRO Learning, (1992).

[4] Load, F. M., Applications of Item Response Theory to Practical Testing Problems. Erlbaum, Hillsdale, NJ ,(1980).

[5] "DIYexamer system", http://speed.cis.nctu.edu.tw/~diy (accessed on Augest 22, 2000).


TABLE 3: DIYexamer questionnaire results: percentage and the number of students in parentheses of each question

| Question | Strongly agree | Agree | No opinion | Disagree | Strongly disagree |
|---|---|---|---|---|---|
| Item DIY is interesting. | 12.3 (7) | 43.9 (25) | 21.1 (12) | 15.8 (9) | 7.0 (4) |
| Item DIY is fanciful. | 19.5 (10) | 49.1 (28) | 21.1 (12) | 10.5 (6) | 1.8 (1) |
| I am curious about the testing result of my DIY item. | 26.3 (15) | 59.6 (34) | 10.5 (6) | 3.5 (2) | 0.0 (0) |
| I learned a lot when creating items. | 12.3 (7) | 47.4 (27) | 22.8 (13) | 17.5 (10) | 0.0 (0) |
| I am curious about the teacher's opinion about my DIY item. | 22.8 (13) | 50.9 (29) | 22.8 (13) | 1.8 (1) | 1.8 (1) |
| I am curious about other students' opinions about my DIY item. | 15.8 (9) | 56.1 (32) | 21.1 (12) | 7.0 (4) | 0.0 (0) |
| I studied harder to prepare item DIY. | 10.5 (6) | 54.4 (31) | 21.1 (12) | 14.0 (8) | 0.0 (0) |
| Judging the difficulties of my DIY items is easy. | 40.4 (23) | 38.6 (22) | 14.0 (8) | 7.0 (4) | 0.0 (0) |
| Judging the fitness of my DIY items is difficult. | 36.8 (21) | 49.1 (28) | 8.8 (5) | 5.3 (3) | 0.0 (0) |
| Item DIY by students comes from the laziness of teachers. | 7.0 (4) | 12.3 (7) | 43.9 (25) | 33.3 (19) | 3.5 (2) |
| If possible, I hope such item DIY mode through the whole course can replace conventional testing. | 1.8 (1) | 10.5 (6) | 35.1 (20) | 38.6 (22) | 14.0 (8) |
| Items generated by students are easier than by the teacher. | 7.0 (4) | 36.8 (21) | 28.1 (16) | 24.6 (14) | 3.5 (2) |
| I knew more about the testing material after item DIY procedure. | 8.8 (5) | 50.9 (29) | 22.8 (13) | 15.8 (9) | 1.8 (1) |