นิพนธ์ต้นฉบับ

A comparative study of two computer software programs for item analysis

Boonnart Laisnitsarekul*

Laisnitsarekul B. A comparative study of two computer software programs for item analysis. Chula Med J 1994 Jan; 38(1) : 23-31

The objectives of this study were to determine and compare the quality of two item analysis computer programs in terms of time used, difficulty index, discrimination index, numbers of good items and reliability of test. A MCQs test, 55 items with 462 students, was calculated by the CTIA and IRT programs. Each program provided essential information such as difficulty index, discrimination index, mean, standard deviation, maximum index, minimum index and reliability. The difficulty index and discrimination index were compared between the two programs by Paired t-test. Each of the two programs required one minutes for data preparation. The time used for processing by CTIA and IRT were 18 and 210 minutes, respectively. The difficulty index, discrimination index, reliability and numbers of good items calculated by CTIA program were equal or higher than for the IRT prograsm. When the Division of Academic Affairs, Faculty of Medicine, Chulalongkorn University considered the time used for processing after receiving the raw data from an optical reader and all of the indices, she decided to use CTIA item analysis program for serving the instructors, beginning in academic year 1993, first semester.

Key words : *Item Analysis, Computer Program, Difficulty Index, Discrimination Index*

Reprint request : Laisnitasarkul B, Medical Education Unit, Faculty of Medicine, Chulalongkorn
Unversity, Bangkok 10330, Thailand.

Received for publication. September 13, 1993.

* Medical Education Unit, Faculty of Medicine,Chulalongkorn University.

บุญนาท ลายสนิทเสรีกุล. การเปรียบเทียบโปรแกรมคอมพิวเตอร์เพื่อการวิเคราะห์ข้อสอบ 2 โปรแกรม. จุฬาลงกรณ์เวชสาร 2537 มกราคม; 38(1) : 23-31

การวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบคุณภาพของโปรแกรมคอมพิวเตอร์ เพื่อการวิเคราะห์ข้อสอบ 2 โปรแกรม ในด้านระยะเวลาการเตรียมข้อมูล เวลาในการทำงานของโปรแกรม ค่าระดับความยากง่ายของข้อสอบ ค่าอำนาจจำแนกของข้อสอบ จำนวนข้อสอบที่ดีและค่าความเที่ยง การดำเนินงานได้ใช้ข้อสอบปรนัยจำนวน 55 ข้อ ที่ใช้สอบนิสิตจำนวน 462 คน นำมาวิเคราะห์ข้อสอบด้วยโปรแกรม CTIA และ IRT. ทุกโปรแกรมจะให้ค่าสำคัญได้แก่ ระดับความยากง่าย อำนาจจำแนก ค่ามัชฌิมเลขคณิต ค่าเบี่ยงเบนมาตรฐาน ค่าสูงสุด ค่าต่ำสุด และค่าความเที่ยง. การเปรียบเทียบค่าระดับความยากง่าย และค่าอำนาจจำแนก ใช้สูตรสถิติการเปรียบเทียบความแตกต่างระหว่างคู่ (Paired t-test) ในการเปรียบเทียบระหว่างโปรแกรม ผลการศึกษาพบว่า ระยะเวลาในการเตรียมข้อมูลก่อนการวิเคราะห์ของโปรแกรม CTIA และ IRT เท่ากับ 1 และ 1 นาทีตามลำดับ ระยะเวลาในการทำงานของโปรแกรม CTIA และ IRT เท่ากับ 18 และ 210 นาทีตามลำดับ ค่าระดับความยากง่าย ค่าอำนาจจำแนก ค่าความเที่ยง และจำนวนข้อสอบที่ดีซึ่งคำนวนโดยโปรแกรม CTIA จะเท่ากับหรือสูงกว่า ค่าที่คำนวณโดยโปรแกรม IRT. ผลจากการศึกษาครั้งนี้ ฝ่ายวิชาการ คณะแพทยศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย จะเลือกใช้โปรแกรมคอมพิวเตอร์เพื่อการวิเคราะห์ข้อสอบ CTIA ให้บริการวิเคราะห์ข้อสอบแก่คณาจารย์ของคณะ ในการสอบไล่ประจำภาคเรียนที่ 1 ปีการศึกษา 2536.

Item Analysis techniques constitute some of the most valuable tools that a classroom teacher can apply in attempting to improve the quality of his tests. Item analyses are conducted for four general purposes: (1) to select the best available items for the final form of a test; (2) to identify any structural or content defects in any of the items; (3) to detect learning difficulties of the class as a whole (identifying general content areas or skills that need to be reviewed by the instructor) and (4) to identify for individual students areas of weakness which may be in need of remediation. There are three main elements involved in performing an item analysis. One is concerned with an examination of the difficulty level of the items. Another element involves determining the discriminating power of each item. The third element involves an examination of the effectiveness of the distractors (alternative answers).[1] The conditions for the application of item analyses are: (1) it applies to relative criteria tests (the procedure leads to a choice of questions that tend to maximize variance and ensure discriminatory ranking); (2) it is applicable only to questions scored dichotomously (1,0); and (3) it should not be applied if the total number of students is very small (a minimum of 20 students could be proposed as a 'pragmatic' criterion).[2]

Medical education in recent years has seen a move towards more objective methods of assessing a student's competence. This is true in both undergraduate and postgraduate spheres and the increasing use being made of multiple choice questions (MCQs) was identified in a recent survey of medical schools in the British Isles. While many formats of MCQs have been described, two types have been more widely used than others. In the 'one-from-five' type of question the student has to choose the one 'best' answer from five possibilities. In the second type of question a common stem is followed by five statements or questions (usually called items), any number of which can be correct. In North America a variation, often known as Type K, is used in which the candidate may be asked to mark 'A' if answers 1,2 and 3 only are correct, 'B' if answers 1 and 3 only are correct, 'C' if answers 2 and 4 only are correct, 'D' if answer 4 only is correct and 'E' if the answers are all correct. This evolved from the 'one-from-five' type of question and the only advantage is that a similar marking technique can be adopted, the correct answer to each questions being represented by a single letter.[3] Both types of MCQs, one best and K type, have been more widely used than others in Thai medical schools too.

At Chulalongkorn University, there are two Computer-based item analysis systems within the Staff Development Unit, Division of Academic Affairs. In 1992, Sukamolson[4] from the Language Institute, Chulalongkorn University, created a program named 'Classical Test Item Analysis (CTIA)'. The program was written in Quick BASIC language for 16-bit or 32-bit microcomputers such as IBM PC/XT, IBM PC/AT, 386, 486 or IBM compatibles. In the same year, Kanjanawasee and Khaimook[5] from the Faculty of Education, Chulalongkorn University and Faculty of Sciences and Technology, Prince of Songkla University created a program named 'Item Response Theory (IRT)'. The program was written in FORTRAN 77 language, and developed for 16-bit microcomputers such as IBM PC/XT, IBM PC/AT or IBM compatibles. The indices and other parameters needed for item analysis were calculated and printed out after completion of data entering depending on the number of items and students. The major problem for item analysis is the data entering step. The time used for data entering by hand for 100 test items with 50 students is about 2-4 hours.[6] In the Faculty of Medicine, Chulalongkorn University, the long time of data entering by hand is a major factor which inhibited the item analysis procedure. To solve this problem, in 1992 the faculty administrator purchased an optical reader for checking the students' answer sheets. It can give the students' score and prepare data for item analysis in 5 minutes for 75 test items with 149 students. The author then become interested studying the quality and usefulness of two item analysis programs when used with the optical reader. The result should be basic information for deciding the choice of an appropriate program in the near future.

Objectives

1. to find item-difficulty index, item-discrimination index, reliability of test, and numbers of good items calculated by two item analysis programs.

2. to compare the difficulty index and discrimination index between the two programs.

3. to find and compare the time used for preparing and processing the data.

4. to choose the best item analysis program based on item-difficulty index, item-discrimination index, reliability, time used for preparing the data, and time used for processing the data.

Definitions

1. **Item Analysis:**[7] Every question (item) is analyzed individually. This item analysis records how many students chose the correct answer, how many chose the other distractors, and how many did not answer the question. The overall student group is divided into high and low performance groups by the computer based on their score in this examination. The proportion of each of these groups choosing each possible answer is determined, revealing in each question whether 'good' students chose the correct answer more frequently than 'bad' students. A question scored correct more frequently by 'bad' than 'good' students should be examined carefully to clarify why 'good' students are not choosing the correct answer. Perhaps the question is out of date, perhaps it can be interpreted in more than one way, or perhaps the teaching differs on what is the correct answer. The computer calculates a discrimination index for each question by comparing the performance of 'good' (high scoring) with 'bad' (low scoring) students on the question. The question should discriminate positively in favor of 'good' students. The

printout can list separately questions which seem too easy or too hard, those which students did not answer those with ineffectual distractors, those with low and negative discrimination indices, and those which seem to have two or more possible correct answers. These lists spur examiners to review the question in light of the programs criticisms.

2. **Difficulty index:**[8] The index for measuring the easiness or difficulty of a test question. It is the percentage (%) of students who have correctly answered the test question; it would seen to be more logical to call it the easiness index. It can vary from 0 to 100%. The following formula is used:

$$\text{Difficulty index} = \frac{H+L}{N} \text{ X } 100$$

where H = number of correct answers in the high group

L = number of correct answers in the low group

N = total number of students in both groups

3. **Discrimination index:**[8] An indicator showing how significantly a question discriminates between 'high' and 'low' students. It varies from -1 to +1. The following formula is used:

$$\text{Discrimination index} = 2 \text{ X } \frac{(H-L)}{N}$$

4. **Good item:** This is based on the indexes obtained. As per a World Health Organization suggestion,[8] a question with a difficulty index lying between 30%-70% is acceptable, it, in that range, the discrimination index is 0.25 or higher.

Materials

1. One IBM PC/AT compatible 16-bit microcomputer.

2. One EPSON LX-86 printer.

3. One OPSCAN Model 5 optical reader.

4. TOOLS: Software for the optical reader.

5. Two Item Analysis Software programs: CTIA (Language Institute) and IRT (Education, Sciences and Technology).

6. Diskettes (5 1/4 inches, Double Sided, Double Density).

7. Word Processing Software programs (QEdit, CU-Writer, WordPerfect)

8. Statistical Software (LOTUS 123, EpiStat)

9. A MCQs test (55 items, 462 students)

Methods

1. The optical reader scanned the students' answer sheets to obtain raw data. After scanning the students' answer sheets, the optical reader is given raw data as shown in Fig.1 and Fig.2.

NAME	TYPE	LEN	START	END
NCS Header	Reserved	40	1	40
ID	Numeric	10	41	50
SEX	Alphabet	1	51	51
ANS	1 digit item	150	52	201
TOTAL	Numeric	5	202	206
CR/LF	Reserved	2	207	208

Figure 1. Data file structure set up for optical scanning (OPSCAN).

```
5 0 0 0 0 0 0 0 0 0 0 1 0 8 1 3 9 3 0 0 1          5 3 2 5   # 0 0 0 1    Y
3 2 3 3 5 4 3 2 2 1 4 5 5 2 4     3 3 4 1 5 1 1 2 3 5 2 2 5 1 3 3 2 5 1 1 4 1 3 5 3 3 4 2 2 1 2 1 2 5 4 1 4 5 4 1
5 0 0 0 0 0 0 0 2 0 0 1 0 8 1 3 9 3 0 0 1          5 3 2 5   # 0 0 0 1    Y                0 0 0 1
5 1 4 4 4 3 2 4 4 3 1 3 3 4 3     1 3 4 5 4 5 1 5 3 1 4 2 5 2 5 3 2 3 5 3 3 5 2 1 3 2 1 5 2 4 3 1 2 1 4 2 1 1 2 1
5 0 0 0 0 0 0 0 3 0 0 1 0 8 1 3 9 3 0 0 1          5 3 2 5   # 0 0 0 1    Y                0 0 0 2
3 2 3 3 5 1 3 3 2 1 3 4 3 3 4     1 3 4 1 2 2 1 5 3 5 4 2 5 4 3 4 2 3 4 5 3 5 2 5 3 3 4 2 2 2 3 3 5 5 4 2 4 3 2 1
5 0 0 0 0 0 0 0 4 0 0 1 0 8 1 3 9 3 0 0 1          5 3 2 5   # 0 0 0 1    Y                0 0 0 3
3 1 3 3 5 5 2 2 3 1 4 4 3 4 4     3 3 4 1 5 5 1 4 3 5 4 2 5 4 4 4 2 3 4 5 3 2 3 5 3 2 4 1 4 4 5 1 3 1 4 1 4 3 1 1
5 0 0 0 0 0 0 0 5 0 0 1 0 8 1 3 9 3 0 0 1          5 3 2 5   # 0 0 0 1    Y                   0 0 0 4
3 1 3 3 5 1 2 2 1 1 1 5 4 2 3     3 3 4 1 5 2 1 5 3 3 2 2 5 4 4 4 2 5 4 5 1 2 3 4 3 2 3 4 4 2 1 1 3 5 4 1 1 1 1 1
5 0 0 0 0 0 0 0 6 0 0 1 0 8 1 3 9 3 0 0 1          5 3 2 5   # 0 0 0 1    Y                0 0 0 5
3 2 3 4 5 5 4 3 2 1 4 3 4 2 3     1 2 4 5 5 1 3 1 3 3 1 2 5 4 5 3 2 3 4 5 3 5 5 5 3 2 1 2 4 2 5 2 3 2 5 1 4 3 1 1
```

Figure 2. A sample of raw data created by OPSCAN.

2. Use the word processing software to prepare the raw data. The raw data was prepared as a data file for the two item analysis programs. For the CTIA and IRT programs, the raw data was prepared as CTIA.DAT and IRT.DAT, respectively. A part of a CTIA.DAT file is shown in Fig.3 and a part of an IRT.DAT file is shown in Fig.4

```
IIIIIIII    AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
KEY         32335432214552433415112352251332511413533422121
0001        51444324431334313454515314252532353352132152431
0002        32335133213433413412215354254342345352533422233
0003        31335522314434433415514354254442345323532414451
0004        31335122111542333415215332254442545123432344211
0005        32345543214342342455131331254532345355532124252
0006        33131312314542433413214311223142315155534424314
0007        32232422214542333515115352224332345355432422211
0008        31231112211354433412514422224522352423232331511
0009        31331532211442433415111554231342335122554222451
0010        32345322211132433412215453233542345352331321211
```

Figure. 3 Format of data structure in a CTIA.DAT file.

```
KEY         32335432214552433415112352251332511413533422121
0001        51444324431334313454515314252532353352132152431
0002        32335133213433413412215354254342345352533422233
0003        31335522314434433415514354254442345323532414451
0004        31335122111542333415215332254442545123432344211
0005        32345543214342342455131331254532345355532124252
0006        33131312314542433413214311223142315155534424314
0007        32232422214542333515115352224332345355432422211
0008        31231112211354433412514422224522352423232331511
0009        31331532211442433415111554231342335122554222451
0010        32345322211132433412215453233542345352331321211
```

Figure. 4 Format of data structure in an IRT.DAT file.

3. Run each item analysis program and check the time used.

4. Count the numbers of good items, and then compare between the two programs by Paired t-test.

Results

1. The times used for preparing the CTIA.DAT and IRT.DAT data files were 1 minute and 1 minute, respectively. The times used for processing by the CTIA and IRT programs were 18 minutes and 210 minutes, respectively.

Table 1. Times used for data preparation and data processing by the CTIA and IRT item analysis programs.

Program	Preparing data	Processing Program
CTIA	1 minute	18 minutes
IRT	1 minute	210 minutes

2. The difficulty index and discrimination index calculated by the CTIA and IRT programs are shown in Table 2 and Table 3. When compared by paired t-test, the difficulty index of CTIA is significantly different from the difficulty index of the IRT program at p<.01. The discrimination indexes between the CTIA and IRT programs were not different.

Table 2. Difficulty index, Mean, Standard Deviation, Maximum index and Minimum index calculated by the CTIA and IRT item analysis programs.

Item No.	CTIA	IRT
1	0.881	0.8599
2	0.271	0.1418
3	0.634	0.5696
4	0.712	0.6613
5	0.639	0.5747
6	0.136	0.02
7	0.113	0.02
8	0.671	0.6129
9	0.81	0.7759
10	0.887	0.8676
11	0.42	0.3175
12	0.539	0.4576
13	0.396	0.2895
14	0.461	0.3659
15	0.439	0.3405
16	0.621	0.5544
17	0.922	0.9083
18	0.801	0.7657
19	0.896	0.8778
20	0.487	0.3965
21	0.18	0.0349
22	0.861	0.837
23	0.113	0.02
24	0.632	0.5671
25	0.487	0.3965
26	0.435	0.3354
27	0.833	0.8039
28	0.294	0.1698
29	0.323	0.203
30	0.348	0.2335
31	0.223	0.0858
32	0.864	0.8396
33	0.277	0.1495
34	0.214	0.0756
35	0.113	0.02

36	0.461	0.3659
37	0.165	0.0171
38	0.429	0.3277
39	0.556	0.478
40	0.749	0.7046
41	0.26	0.1291
42	0.452	0.3557
43	0.353	0.2386
44	0.42	0.3175
45	0.277	0.1495
46	0.177	0.0323
47	0.42	0.3175
48	0.219	0.0807
49	0.199	0.0578
50	0.82	0.7886
51	0.738	0.6919
52	0.37	0.259
53	0.387	0.2793
54	0.045	0.02
55	0.885	0.865
MEAN	0.478454	0.393178
S.D.	0.253138	0.288302
MAX	0.922	0.9083
MIN	0.045	0.0171

Table 3. Discrimination index, Mean, Standard Deviation, Maximum index and Minimum index calculated by the CTIA and IRT item analysis programs.

Item No.	CTIA	IRT
1	0.153	0.2824
2	- 0.008	0.0057
3	0.29	0.2285
4	0.363	0.3559
5	0.355	0.3305
6	0.024	0.0161
7	0.04	0.0966
8	0.435	0.3986
9	0.298	0.4396
10	0.194	0.4457
11	0.444	0.3677
12	0.444	0.3853
13	0.194	0.1818
14	0.226	0.1784
15	0.226	0.2315
16	0.363	0.3408
17	0.218	0.5212
18	0.226	0.2981
19	0.315	0.68
20	0.371	0.2697
21	0.04	0.0589

22	0.234	0.4343
23	0.073	0.123
24	0.339	0.3378
25	0.516	0.4501
26	0.298	0.2689
27	0.387	0.5474
28	0.25	0.2458
29	0.331	0.2685
30	0.347	0.3028
31	0.024	0.0413
32	0.266	0.4853
33	0.282	0.2477
34	0.129	0.1018
35	-0.04	-0.0742
36	0.46	0.3689
37	0.129	0.1317
38	0.25	0.2732
39	0.234	0.2509
40	0.177	0.2355
41	0.185	0.1143
42	0.371	0.3166
43	0.355	0.2957
44	0.548	0.4327
45	0.048	0.0697
46	0.048	0.0458
47	0.492	0.3878
48	0.113	0.0952
49	0.202	0.1897
50	0.395	0.5403
51	0.379	0.4301
52	0.331	0.2556
53	0.5	0.4169
54	-0.04	-0.1328
55	0.202	0.3637
MEAN	0.255018	0.272065
S.D.	0.149393	0.165607
MAX	0.548	0.68
MIN	-0.04	-0.1328

3. The reliability of test calculated by the CTIA and IRT programs were 0.70 and 0.69, respectively. When counting the number of good items, there were 36.36% for CTIA and 32.73% for the IRT program.

Table 4. Reliability and numbers of good items calculated by the CTIA and IRT programs.

	CTIA	IRT
1. Reliability	0.702	0.6905
2. Number of good items	20 in 55 items (36.36%)	18 in 55 items (32.73%)

Discussion

In the past at the Faculty of Medicine, Chulalongkorn University, instructors entered raw data by keyboard when they needed to do item analysis. The time required for 100 items with 50 students was 2-4 hours, depended on the instructors' experience.[6] The time used for entering raw data can be greatly decreased when the optical reader is used. When using the OPSCAN Model 5, the time used for 100 items with 150 students was only 5 minutes.

The Division of Academic Affairs, Faculty of Medicine, Chulalongkorn University, would like to serve faculty staff with computer-based item analysis. The program selected should be suitable for use with the Faculty's optical reader and the time required for all processing should be quite low. From th results, of this research it appears that the CTIA item analysis program is more appropriate than the IRT program. The time used for preparing data and processing program by CTIA was less than for the IRT program. The difficulty index, discrimination index, reliability, and numbers of good items calculated by the CTIA program are equal or higher than for the IRT program. When using the World Health Organization's criteria to assess the difficulty index and discrimination index, it is shown the CTIA provided a higher number of good items than IRT. Hubbard and Clemans (1961)[9], Schumacher (1971)[10] and Cox and Ewan (1988)[11] suggested that a good test should have reliability at 0.70 or higher. The reliability calculated by CTIA in our research was 0.70, and this indicates that the test is acceptable. On the other hand, the reliability index by the IRT program was 0.69, which indicates a poor test. The IRT program was written in FORTRAN language which is appropriate for mainframe computers more than for microcomputers. After comparing the results, the Division of Academic Affairs, Faculty of Medicine, Chulalongkorn University, decided to use the CTIA item program for serving the faculty instructors.

Summary

The objectives of this study were to determine and compare the quality of two item analysis computer programs in terms of time used, difficulty index, discrimination index, numbers of good items and the reliability of the test. A MCQs test, 55 items with 462 students, was calculated by use of the CTIA and IRT programs. Each program provided essential information such as the difficulty index, discrimination index, mean, standard deviation, maximum index, minimum index and reliability. The difficulty index and discrimination index were compared between the two programs by Paired t-test. The times used for preparing data files by CTIA and IRT were each 1 minute. The times used for data processing by the CTIA and IRT were 18 and 210 minutes, respectively. The difficulty index, discrimination index, reliability and numbers of good items calculated by the CTIA program were equal or higher than for IRT. When the Division of Academic Affairs, Faculty of Medicine, Chulalongkorn University considered the time used for all processing after receiving the raw data from the optical reader, and all of indices, she decided to use the CTIA item analysis program for serving the instructors in the 1993 academic, first semester.

Acknowledgement

The author is grateful to Dr. Supath Sukamolson, Dr.Sirichai Kanjanawasee, Dr. Knit Khaimook and personnel of the Staff Development Unit, Division of Academic Affairs, Chulalongkorn University Presidents Office for their contribution towards the CTIA and IRT item analysis programs. I am also indebted to Assistant Professor Dr. Pornarong Chotiwan, Head of Computer Unit, Faculty of Medicine, Chulalongkorn University who advised on use of the optical reader.

References

1. Payne DA. The Specification and Measurement of Learning Outcomes. Toronto: Xerox College Publishing, 1968: 145
2. Guilbert JJ. Educational Handbook for Health Personnel. Rev ed. Geneva: World Health Organization, 1987: 4.72
3. Harden RMcG. Constructing Multiple Choice Questions of the Multiple True/False Type. ASME Medical Education Booklet Number 10. Dundee: Association for the Study of Medical Education, 1979: 2
4. Sukamolson S. Classical Test Item Analysis and Grading Manual(V.6.30). Bangkok: Chulalongkorn University Language Institute, 1992.
5. Kanjanawasee S, Khaimook K. Item Response Theory: Bayesian Procedure Manual (V.1.0). Bangkok: Faculty of Education, Chulalongkorn University, 1992.
6. Chongtrakul P, Jaroongdaechakul M. Microcomputer program for item analysis. Chula Med J 1988 Feb; 32(2):199-206
7. Cox K, Ewan CE. The Medical Teacher. 2nd ed. London: Churchill Livingstone, 1988: 164
8. Guilbert JJ. Educational Handbook for Health Personnel. Revised ed. Geneva: World Health Organization, 1987. 4.68-4.69
9. Hubbard JP, Clemans WV. Multiple-Choice Examinations in Medicine: A Guide for Examiner and Examinee. Philadelphia: Lea & Febiger, 1961: 71
10. Schumacher CF. Scoring and analysis. In: Hubbard JP, ed. Measuring Medical Education. Philadelphia: Lea & Febiger,1971.
11. Cox K, Ewan CE. The Medical Teacher. 2nd.ed. London:Churchill Livingstone, 1988: 163