UDC 159.9.072.5

紙筆テストとコンピューターベースト・テストの解答所要時間と得点分布の比較可能性

藤芳　衛*
藤芳　明生**
石塚　智一*（2000）

要　約

本研究は通常の紙筆テストと 1995 年に開発したコンピューターベースト・テストの比較可能性を評価するため，大学入試センター試験の国語・数学・英語の 3 教科の解答所要時間と得点分布に関して両テスト・メディアを比較する実験的分析である。実験に使用したコンピューターベースト・テストはペン・コンピュータを使用し，紙筆テストをできる限り忠実にシミュレートできるように設計した。紙筆テストの紙をコンピュータ画面に，鉛筆を電子ペンにそのまま置き換えた。

両テスト・メディアの解答所要時間と得点分布の比較可能性を検討するため行った 1996 年の先行実験において，国語の得点だけに有意差が認められた。しかし，紙筆テスト群とコンピューターベースト・テスト群とは被験者群が異なるため，有意差の原因がテスト・メディアの違いに起因するか，あるいは被験者群の相違に依存するかが明らかではなかった。

本評価実験では実験計画としてラテン方格法を採用し，両テスト・メディアのテストを同一の被験者群に同時に実施することにより，被験者群の相違に依存する要因の影響をできる限り排除した実験を行った。この結果，大学入試センター試験の国語・数学・英語の 3 教科に関して通常の紙筆テストと本コンピューターベースト・テストで測定した解答所要時間と得点分布は比較可能であった。

* 大学入試センター研究開発部
** 茨城大学工学部

Comparability of Paper-and-Pencil Tests and Computer-Based Tests in Terms of Distributions of Completion Time and Score

Mamoru Fujiyoshi*
Akio Fujiyoshi**
Tomoichi Ishizuka*

Abstract

This experimental study was conducted to assess the comparability of conventional paper-and-pencil tests and the computer-based tests which were developed in 1995. The assessment was to compare the distributions of completion time (time needed to complete test items) and score on the two test media used for Japanese, Mathematics and English tests in the National Center Test for University Admissions. The computer-based test was designed to simulate the conventional paper-and-pencil test as faithfully as possible. It employed a pen computer. A paper-and-pencil test sheet was displayed on the computer screen, with the pencil being replaced by the electronic pen.

The comparability of distributions of completion time and score was experimentally assessed in 1996. As the subject groups to whom administered the paper-and-pencil and computer-based tests were administered, were different, it was unclear as to whether the cause of the significant difference noted in scores for Japanese was due to the difference in the test media, or to the difference between the two subject groups.

This experimental assessment employed a Latin-square design, and both subject groups were simultaneously administered tests on the two test media to minimize the effects of difference between subject groups. It was concluded that the distributions of completion time and score for the paper-and-pencil tests and the computer-based tests were comparable.

Key Words : paper-and-pencil test, computer-based test, item cumulative time-completion rate curve, item cumulative time-score rate curve, National Center Test for University Admissions

1 Introduction

This experimental assessment was conducted for the purposes of investigating the comparability of distributions of completion time (time needed to complete test items) and score measured by the conventional paper-and-pencil test (PPT) and the computer-based test (CBT) developed in 1995. Employing a pen computer, the CBT was designed to simulate the answering process of the PPT of the National Center Test for University Admissions using optical readable marking sheets as faithfully as possible (Fujiyoshi & Ishizuka, 1996). The PPT sheet was displayed on the computer screen, with the pencil being replaced by the electronic pen.

The results of the experimental assessment in 1996 suggested that we should improve the experimental methodology needed improvement. The distributions of completion time and score for

* Research Division, The National Center for University Entrance Examinations
** Faculty of Engineering, Ibaraki University

Japanese, Mathematics, and English tests in the National Center Test for University Admissions differed minimally between the PPT and the CBT, however it was only for Japanese tests that the mean score for the CBT was significantly lower than that for the PPT (Fujiyoshi & Ishizuka, 1996). Since the PPT subject group and the CBT subject group were different, it was unclear as to whether the cause of this significant difference was related to the difference in the two test media, or to the difference between the two subject groups.

This new experimental assessment employed a Latin-square design, that is, both subject groups were simultaneously administered the tests on the two test media to eliminate the effects of difference between subject groups, in order to figure out the problems in the previous assessment. The Latin-square design is considered to be one of the most precise experimental designs to detect effects of test media. Mazzeo et al. used a single-group counterbalanced equating design and investigated the comparability of scores from the PPTs and the CBTs of the College-Level Examination Program (CLEP) General Examinations in Mathematics and English Composition (Mazzeo, Druesne, Raffeld, Checketts, & Muhlstein, 1992). The single-group counterbalanced equating design is the same as the design II of the six designs for test equating proposed by Angoff (1971). In general, for a given sample size, greater precision is obtained from this design than from random groups or anchor-test designs (Mazzeo et al, 1992). The Latin-square design is similar to the single-group counterbalanced equating design. For the single-group counterbalanced equating design, the difficulty of test sets should be equated in advance, However, it is not necessary for the Latin-square design.

In order to compare the distributions of completion time and test score for the PPTs and the CBTs, item cumulative time-completion rate curves and item cumulative time-score rate curves were newly defined (see Appendix) and employed in this study.

2 The CBT

The CBT was designed to simulate the PPTs by employing the optical readable marking sheets as used in the National Center Test as faithfully as possible. Pen computers (Amity SV) supplied by Mitsubishi Electric Corporation were employed in the research. The computers are of light-weight design, 29.6 cm long, 22.8 cm wide and 2.54 cm thick. The upper face of the computer is equipped with a 16-gray scale monochrome liquid crystal display, which is 20.6 cm long and 15.5 cm wide. Resolution of the display is 768×1024 pixels.

An electronic pen is used as an input device. The position of the tip of the pen and the status of the switches on its tip and side, are detected whenever the pen is touched on the display. Data may therefore be entered by simply touching the pen on the display.

Software was written in Visual BASIC Ver. 2.0 (Microsoft Corporation), and run under MS Windows Ver. 3.1 (Microsoft Corporation). Figure 1 shows the first question screen of the Mathematics test. Notes have been handwritten on the screen.

The procedure of the CBT is similar to the PPT in that all questions may be answered using a single electronic pen. Questions on any page may be displayed on the screen as required, as well as any notes handwritten on the screen, and answers can be entered in the electronic marking sheet field on the screen—all functions performed with the electronic pen. Questions may be reviewed as often as necessary, and answers can be corrected as required.

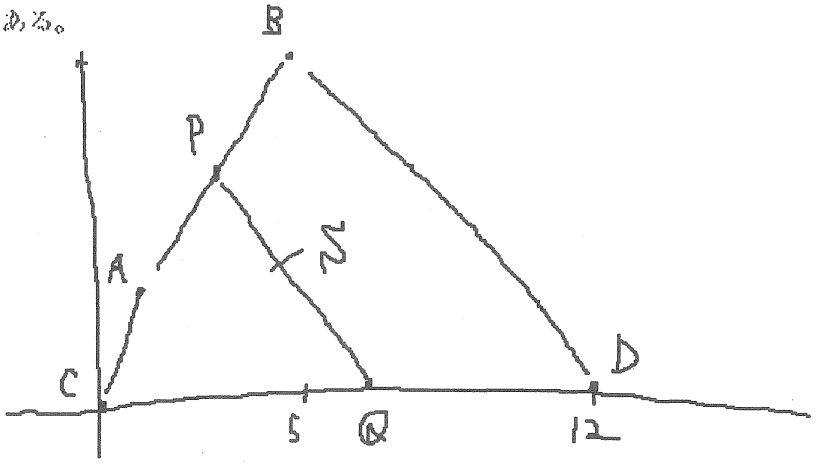
Recording of the answering process is fully automated. Each time a page is turned over, the page number and time are recorded automatically on the internal hard disk drive. Each time the electronic pen is touched on a marking sheet field the item number, choice number and time are recorded. The score is also recorded if the answer is correct.

1 (1) | 1 (2) | 1 (3) | **2** | 3

4点 A(1, 3), B(5, 8), C(0, 0), D(12, 0)がある。線分 AB 上の動点 P と線分 CD 上の動点 Q に対し，線分 PQ の中点を S とする。このとき点 S の存在する範囲は，4点

K($\frac{5}{2}$, [アイ]), L($\frac{1}{2}$, $\frac{[ウエ]}{[オ]}$), M($\frac{[カキ]}{[ク]}$, $\frac{3}{2}$), N([ケコ], [サ])

を頂点とする四角形の周および内部である。そして四角形 KLMN の面積は [シス] である。



中断

終了

	解答欄												
	−	±	0	1	2	3	4	5	6	7	8	9	＊
ア	㊀	㊁	⓪	①	②	③	●	⑤	⑥	⑦	⑧	⑨	⊛
イ	㊀	㊁	⓪	①	②	③	④	⑤	⑥	⑦	⑧	⑨	●
ウ	㊀	㊁	⓪	①	②	●	④	⑤	⑥	⑦	⑧	⑨	⊛
エ	㊀	㊁	⓪	①	②	③	④	⑤	⑥	⑦	⑧	⑨	●
オ	㊀	㊁	⓪	①	●	③	④	⑤	⑥	⑦	⑧	⑨	⊛
カ	㊀	㊁	⓪	●	②	③	④	⑤	⑥	⑦	⑧	⑨	⊛
キ	㊀	㊁	⓪	①	②	●	④	⑤	⑥	⑦	⑧	⑨	⊛
ク	㊀	㊁	⓪	①	●	③	④	⑤	⑥	⑦	⑧	⑨	⊛

	解答欄												
	−	±	0	1	2	3	4	5	6	7	8	9	＊
ケ	㊀	㊁	⓪	●	②	③	④	⑤	⑥	⑦	⑧	⑨	⊛
コ	㊀	㊁	⓪	①	②	③	④	⑤	⑥	●	⑧	⑨	⊛
サ	㊀	㊁	⓪	①	●	③	④	⑤	⑥	⑦	⑧	⑨	⊛
シ	㊀	㊁	⓪	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⊛
ス	㊀	㊁	⓪	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⊛

Figure 1

The Pen Computer Screen of the Computer-Based Test for Mathematics t-Test

3 Experimental Assessment

3.1 Purpose

The purpose of this assessment was to investigate the comparability of the conventional PPTs and the CBTs in terms of the distributions of completion time and score in the National Center Test for University Admissions. The assessment used the Latin-square design to resolve problems of methodology revealed in the previous assessment in 1996 (Fujiyoshi & Ishizuka, 1996).

3.2 Method

Due to the nature of the tests it is necessary to preclude the same subjects being administered the same questions on both of the two test media. A 2×2 Latin-square design using 16 repeats was employed. Table 1 shows the design plan.

Table 1

The Experimental Plan of Latin-Square Design

		Subject Group	
		Group 1	Group 2
Test Media	Computer-Based Test	Test Set A	Test Set B
	Paper-and-Pencil Test	Test Set B	Test Set A

The factors in the experiment were the two levels of test media (PPT and CBT), the two levels of subject groups (group 1 and group 2), and the two levels of test sets (set A and set B).

The test media factor was the two levels of the PPT and the CBT. The questions for the PPT were presented in a conventional booklet format. In that the use of video recording method required considerable time and effort in reading the test data (Fujiyoshi & Ishizuka, 1996), answers were recorded automatically using the electronic marking sheet system developed from the electronic marking sheet portion of the CBT.

Subjects were 32 first year university students, entering university in 1996, and having taken the National Center Test for University Admissions in Japanese, Mathematics and English.

The test sets used in the experiment consisted of set A and set B for each of the Japanese, Mathematics and English exams, and were prepared from questions previously used in the National Center Test. The number of questions in each test set was approximately equivalent to the amount of material for 40 minutes of the National Center Test.

Each cell of the Latin-square design contained two additional factors in the experiment. The university department factor consisted of two levels, literature and natural sciences, and the administration order factor consisted of the two levels, PPT-first and CBT-first administrations. Male and female subjects were approximately equal in number.

The test procedure started with issuing instructions, after which the subjects answered test questions in accordance with the work-limit method. A time limit was not required.

3.3 Results

3.3.1 ANOVA of Effects of Test Media on the Completion Time and the Test Scores

The three factors in the Latin-square design for the experiment were analyzed by ANOVA to determine their effects on completion time and score. The results of ANOVA for completion time of the tests are shown in Table 2. The results of ANOVA for test scores are shown in Table 3.

The distributions of completion time for the PPTs and the CBTs were approximately the same for Japanese, Mathematics and English. Box-and-whisker plots representing distribution of completion time of the tests are shown in Figure 2. The '+' symbols shown in the box-and-whisker plots represent the mean of a distribution.

Table 2

The Result of ANOVA for the Effects of Test Media, Subject Groups and Test Sets on the Completion Time of Tests

Subject Area	Source	DF	Sum of Squares	Mean Square	F Value	p<
	Test Media	1	2.24	2.24	0.04	
	Subject Group	1	20.95	20.95	0.35	
Japanese	Test Set	1	22.33	22.33	0.38	
	Error	60	3551.61	59.19		
	Corrected Total	63	3597.13			
	Test Media	1	14.29	14.29	0.13	
	Subject Group	1	301.02	301.02	2.77	
Mathematics	Test Set	1	2224.07	2224.07	20.43	.001
	Error	60	6531.54	108.86		
	Corrected Total	63	9070.92			
	Test Media	1	129.11	129.11	1.32	
	Subject Group	1	5.48	5.48	0.06	
English	Test Set	1	1003.62	1003.62	10.28	.01
	Error	60	5859.93	97.67		
	Corrected Total	63	6998.13			

Table3

The Result of ANOVA for the Effects of Test Media, Subject Groups and Test Sets on the Test Scores

Subject Area	Source	DF	Sum of Squares	Mean Square	F Value	p<
	Test Media	1	121.76	121.76	0.74	
	Subject Group	1	313.33	313.33	1.91	
Japanese	Test Set	1	90.12	90.12	0.55	
	Error	60	9851.37	164.19		
	Corrected Total	63	10376.59			
	Test Media	1	6.10	6.10	0.01	
	Subject Group	1	23.47	23.41	0.05	
Mathematics	Test Set	1	4281.77	4281.77	9.92	
	Error	60	25910.81	431.85		
	Corrected Total	63	30222.09			
	Test Media	1	965.80	965.80	6.20	.05
	Subject Group	1	776.12	776.12	4.98	.05
English	Test Set	1	2.60	2.60	0.02	
	Error	60	9349.37	155.82		
	Corrected Total	63	11093.88			

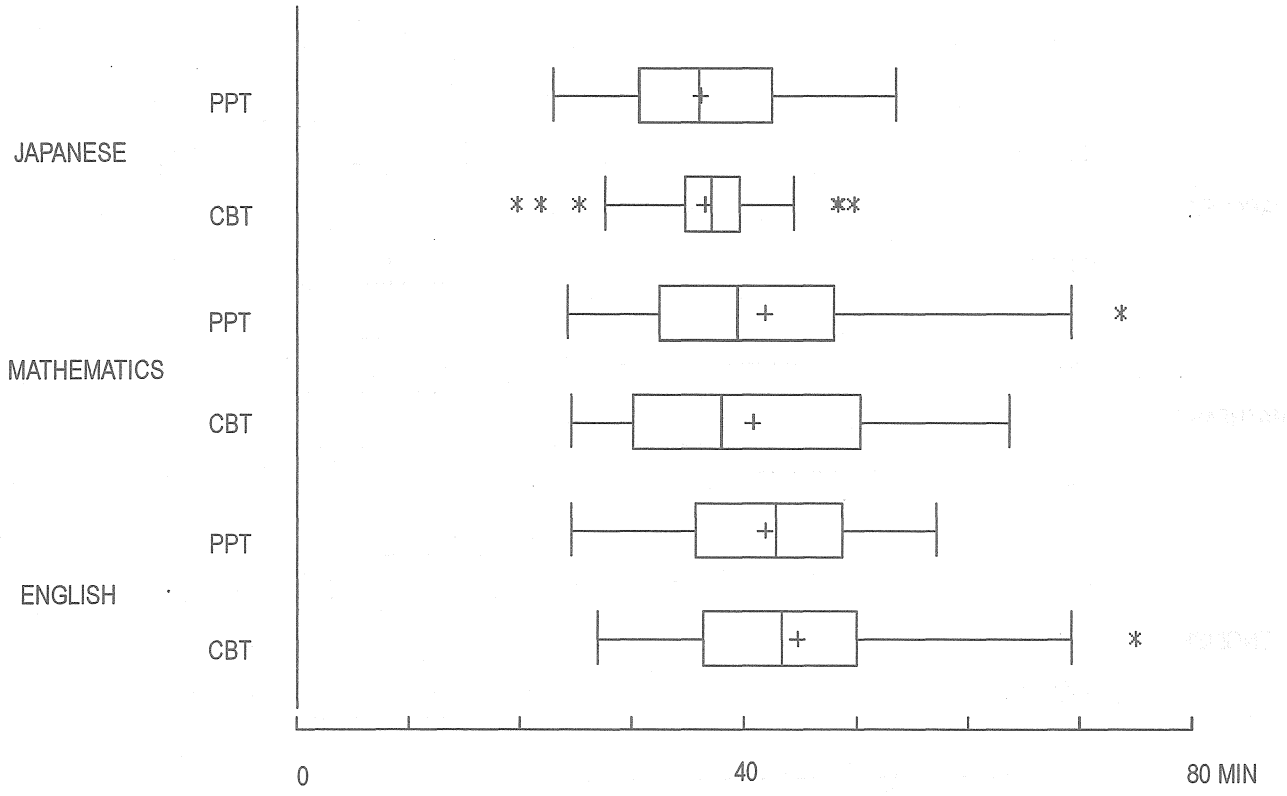


Figure 2
The Box-and-Whisker Plots of the Distributions of Completion Time

As shown in Table 2, the results of ANOVA for completion time indicate that the main effects of the test media factor on completion time were not significant for all three subject areas. Furthermore, the main effects of the subject group factor were not significant. However, the main effects of the test set factor were recognized as significant for Mathematics and English.

The distributions of score for the PPTs and the CBTs were approximately the same for Japanese and Mathematics, however, for English the score distribution for the CBT was slightly higher than that for the PPT. Box-and-whisker plots representing score distributions for the three subject areas are shown in Figure 3.

As shown in Table 3, the results of ANOVA for test scores indicate that the main effect of the test media factor on score was not significant for Japanese and Mathematics but was for English. As with the test media factor, the main effect of the subject group factor was significant only for English. The main effect of the test set factor was not significant for Japanese and English but was for Mathematics.

3.3.2 The Results of t-Test for Order of Administration

The paired t-tests were conducted to investigate the effects of order of administration of the tests on the completion time and score. The results of t-test for distributions of completion time for the three subject areas are shown in Table 4. The results of t-test for score distributions for the three subject areas are shown in Table 5. The total group of 32 subjects took the first test, followed by the second test. Half of the subjects were assigned to the CBT-first group, and were administered the first test as a CBT, then

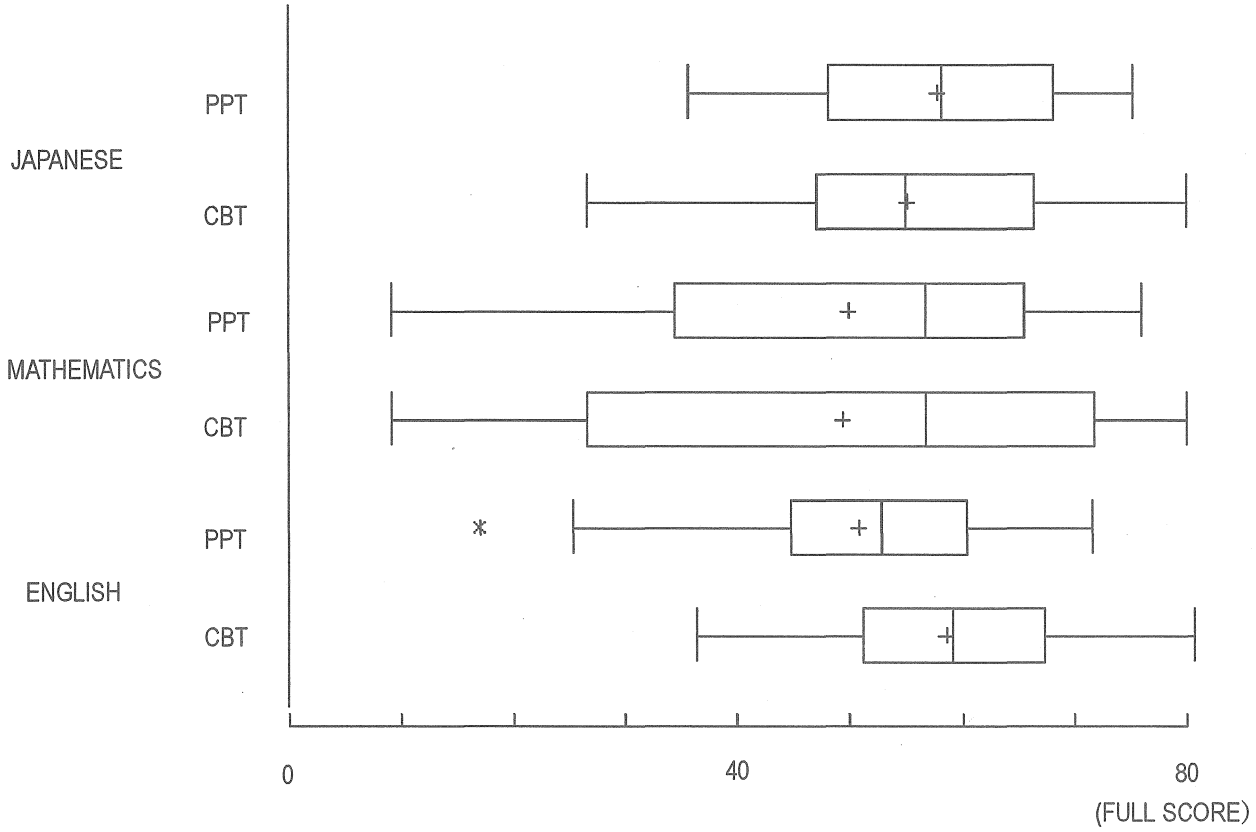


Figure 3

The Box-and-Whisker Plots of the Distributions of Score

were administered the second test as a PPT. The remaining 16 subjects were assigned to the PPT-first group, being given administered the first test as a PPT, then were administered the second test as a CBT. Table 4 shows the mean and standard deviation of the completion time for the first test, those for the second test, those for the differences of the first and second test, the t value, and the level of significance. Each row is partitioned into the total group, the PPT-first group, and the CBT-first group. Table 5 shows the mean and standard deviation of score for the first test, those for the second test, those for the differences of the first and second test, the t value, and the level of significance. Each row is partitioned into the total group, the PPT-first group, and the CBT-first group.

For the distributions of completion time, there was a common tendency that the means for completion time for the second test were relatively smaller than those for the first test in all three subject areas for the total group, the CBT-first group and the PPT-first group. Especially, there was a significant difference in the subject area of Japanese for the total group.

The mean of completion time for the total group was slightly larger for the first test than for the second test for all three subject areas. The means of the differences in completion time were 4.40 minutes for Japanese, 3.02 minutes for Mathematics and 3.70 minutes for English. The results of t-test showed no significant differences for Mathematics and English, but there was a significant difference for Japanese.

The means of completion time for the CBT-first group were the same as those of the total group. The

Table 4

The Result of t-Test for the Effects of Administration Order of the Tests on the Distributions of Completion Time

Subject Area		Japanese			Mathematics			English		
		Total Group	PPT-First Group	CBT-First Group	Total Group	PPT-First Group	CBT-First Group	Total Group	PPT-First Group	CBT-First Group
Number		32	16	16	32	16	16	32	16	16
First Test	Mean	38.75	37.38	40.12	43.00	42.38	43.63	45.26	46.19	44.33
	SD	7.98	7.57	8.38	13.33	13.23	13.83	11.24	13.14	9.31
Second Test	Mean	34.35	32.61	36.09	39.99	40.31	39.66	41.56	39.65	43.47
	SD	6.51	6.35	6.40	10.50	10.72	10.62	9.61	8.81	10.27
Difference	Mean	4.40	4.78	4.03	3.02	2.07	3.96	3.70	6.54	0.86
	SD	9.45	10.53	8.57	20.72	20.13	21.91	14.75	13.15	16.12
t value		2.63	1.81	1.88	0.82	0.41	0.72	1.42	1.99	0.21
p<		.05								

Table 5

The Result of t-Test for the Effects of Administration Order of the Tests on the Distributions of Score

Subject Area		Japanese			Mathematics			English		
		Total Group	PPT-First Group	CBT-First Group	Total Group	PPT-First Group	CBT-First Group	Total Group	PPT-First Group	CBT-First Group
Number		32	16	16	32	16	16	32	16	16
First Test	Mean	58.46	58.30	58.62	50.64	44.93	56.35	54.03	56.99	51.08
	SD	11.70	6.35	10.98	22.03	24.14	18.73	12.00	13.31	10.10
Second Test	Mean	54.66	57.26	52.06	48.99	43.90	54.08	55.68	50.86	60.50
	SD	13.80	13.09	14.42	22.10	21.09	22.57	14.57	16.29	11.14
Difference	Mean	3.80	1.04	6.56	1.65	1.03	2.26	-1.65	6.12	-9.42
	SD	11.90	10.59	12.82	23.55	21.32	26.28	14.50	13.84	10.69
t value		1.81	0.39	2.05	0.40	0.19	0.34	-0.64	1.77	-3.52
p<										.01

means were slightly larger for the CBT (first test) than for the PPT (second test) for all three subject areas. The means of the differences of completion time were 4.78 minutes for Japanese, 2.07 minutes for Mathematics and 6.54 minutes for English. The results of t-test showed no significant differences for all three subject areas, however, the t values for Japanese and English were considerably high.

The means of completion time for the PPT-first group were slightly larger for the PPT (first test) than for the CBT (second test) for all three subject areas. This tendency is similar to that of the total group and the CBT-first group. The means of the differences in completion time were 4.03 minutes for Japanese, 3.96 minutes for Mathematics and 0.86 minutes for English. The results of the t-test showed no significant differences for all three subject areas.

On the other hand, for the distributions of score,

there was a tendency that the mean of score for the second test was approximately the same as that for the first test in Japanese and Mathematics, and lower than that for the first test in English.

The means of score for the first test and the second test were almost the same for all three subject areas for the total group. The means of the differences in score were 3.80 points for Japanese, 1.65 points for Mathematics and -1.65 points for English. The results of the t-test showed no significant differences for all three subject areas.

For the CBT-first group, the mean of score for the CBT was similar to that of the PPT in Japanese and Mathematics, and was considerably higher than that of PPT in English. The means for the differences in score were 1.04 points for Japanese, 1.03 points for Mathematics and 6.12 points for English. The results of the t-test showed no significant differences for all three subject areas.

For the PPT-first group, The means of score for the PPT (first test) were slightly higher than those for the CBT (second test) in Japanese and Mathematics, whereas the mean of score for the CBT (second test) in English was significantly higher. The means of the differences in score were 6.56 points for Japanese, 2.26 points for Mathematics and -9.42 points for English.

The distribution of score for the CBT in English was considerably higher than that for the PPT, irrespective of the order of administration. A significant difference was noted in the distributions of score for the PPT-first group.

3.3.3 Comparisons Using Item Cumulative Time-Completion Rate Curves

Item cumulative time-completion rate curves were newly developed to provide more detailed comparisons of distributions of completion time for the PPTs and the CBTs. An item cumulative time-completion rate curve is a set of points on a coordinate system, with time needed to complete the items on the horizontal axis and the relative cumulative frequency of items answered within the time on the vertical axis (see Appendix for the detailed definition). The item cumulative time-completion rate curves for the two test media for Japanese, Mathematics and English are shown in Figure 4. The bold lines relate to PPTs, and the thin lines relate to the CBTs.

The time-completion rate curves for both test media approximated each other for all three subject areas. The curve for the PPT in Japanese is slightly higher than that for the CBT, however the two curves are matched over most of the range. In contrast, the curve for the CBT in Mathematics is higher than that for the PPT, but the two curves are parallel over most of the range. The two curves for English are very closely matched, where curves for the CBT is overlapped by the curve for the PPT.

The item cumulative time-completion rate curves were plotted for a large number of points and consequently are very smooth. Table 6 shows the total number of points plotted from the test data of 32 subjects and the mean number of points for a subject on both curves.

Table 6

The Total Number of Points and the Mean Number of Points for One Subject on the Item Cumulative Time-Completion Rate Curves

Subject Area	Paper-and-Pencil Test		Computer-Based Test	
	Total Number of Points	Mean Number of Points for One Subject	Total Number of Points	Mean Number of Points for One Subject
Japanese	1220	38.1	1220	38.1
Mathematics	3542	110.7	3753	117.3
English	6217	194.3	6293	196.7

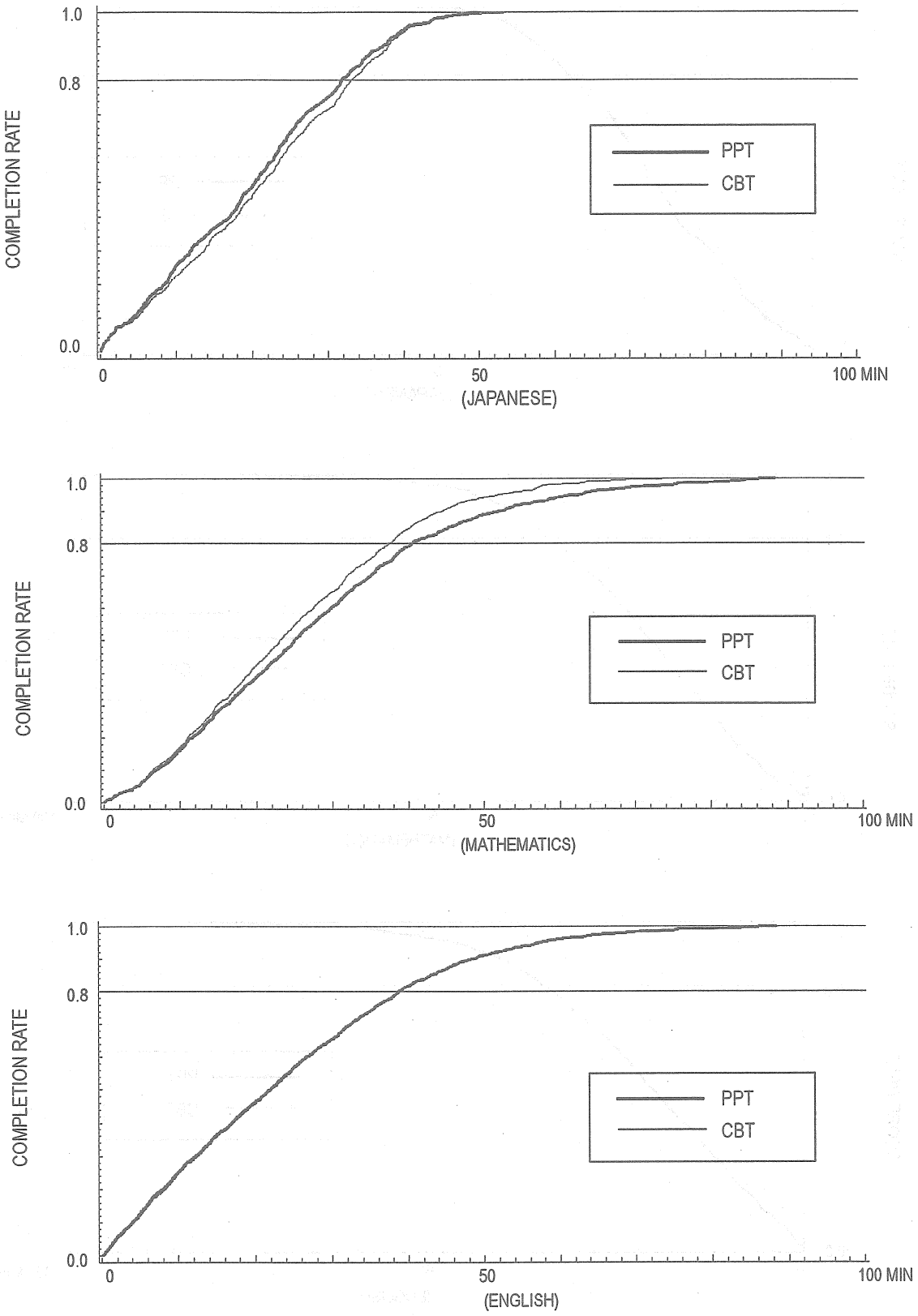


Figure 4

The Item Cumulative Time-Completion Rate Curves for thePaper-and-Pencil and Computer-Based Tests

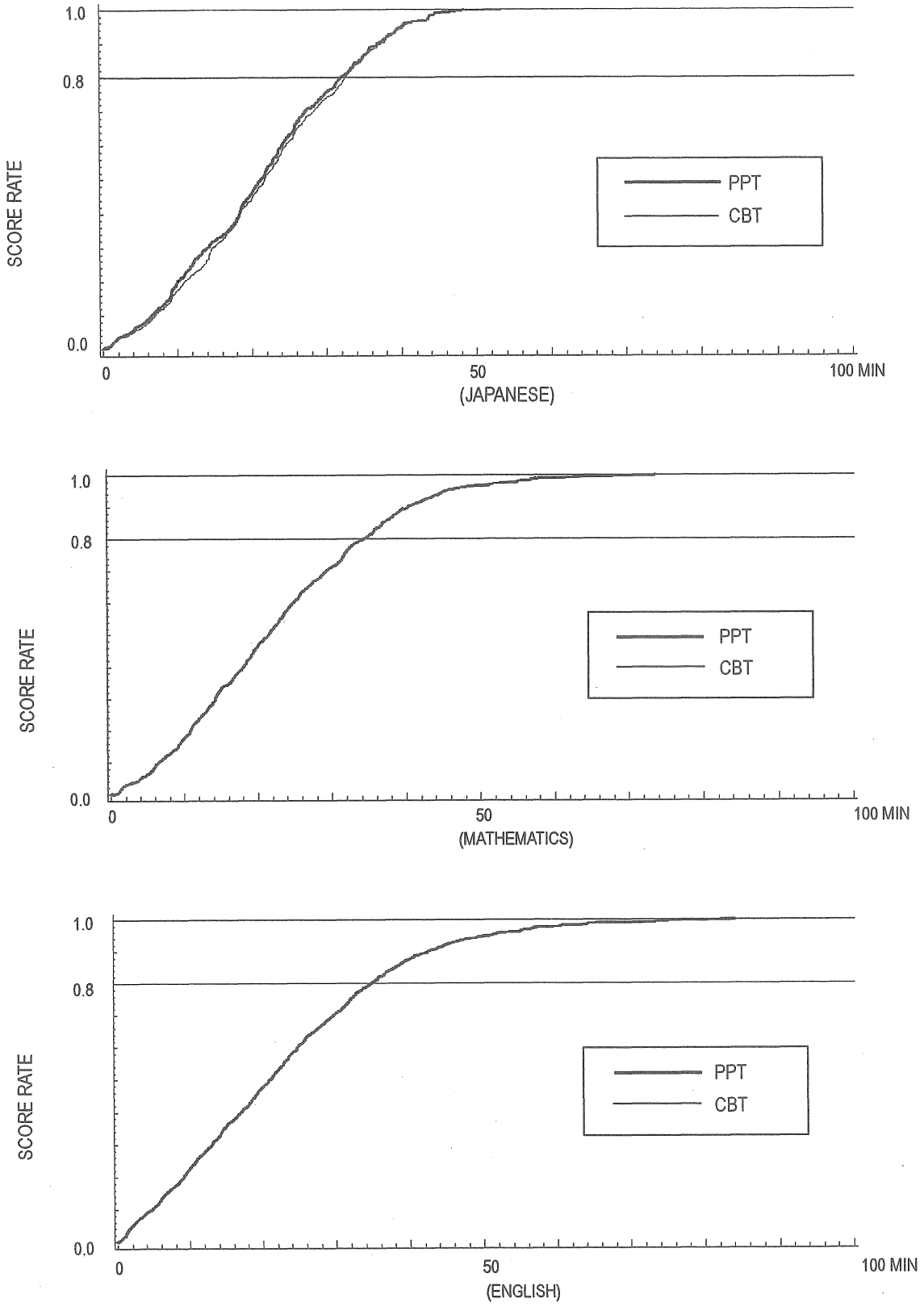


Figure 5

The Item Cumulative Time-Score Rate Curves for the Paper-and-Pencil and Computer-Based Tests

3.3.4 Comparisons Using Item Cumulative Time-score Rate Curves

Item cumulative time-score rate curves were newly developed to provide a more detailed comparison of score distributions for the PPTs and the CBTs. An item cumulative time-score rate curve is a set of points on coordinate system with time needed to complete an item on the horizontal axis and the relative cumulative score of items answered within the time on the vertical axis (see Appendix for the detailed definition). The item cumulative time-score rate curves for the two test media for the three subject areas are shown in Figure 5. The bold lines relate to the PPTs, and the thin lines relate to the CBTs.

The item cumulative time-score rate curves for the PPT and the CBT match well for Japanese, Mathematics and English. Although a significant difference was apparent between the both means of score for English, the item cumulative time-score rate curves for the PPT and CBT are collinear over most of the range.

The item cumulative time-score rate curves were plotted for a large number of points and are consequently very smooth. Table 7 shows the total number of points plotted from the test data for 32 subjects and the mean number of points for a subject on both curves.

Table 7

The Total Number of Points and the Mean Number of Points for One Subject on the Item Cumulative Time-Score Rate Curves

Subject Area	Paper-and-Pencil Test		Computer-Based Test	
	Total Number of Points	Mean Number of Points for One Subject	Total Number of Points	Mean Number of Points for One Subject
Japanese	777	24.3	742	23.2
Mathematics	1474	46.1	1520	47.5
English	2676	83.6	2751	86.0

3.4 Discussion

It was found in the results of this experimental study that the distributions of completion time and score of the PPT and the CBT generally approximated each other for Japanese, Mathematics and English tests in the National Center Test for University Admissions but the English score distribution for the CBTs was slightly higher than that for the PPTs. In the results of ANOVA, except for English score, no significant main effects due to test media were apparent on completion time and score for the three subject areas (see Table 2 and Table 3). The item cumulative time-completion rate curves for the PPT and the CBT approximated each other for all three subject areas (see Figure 4). The item cumulative time-score rate curves were also very well matched for all three subject areas (see Figure 5). Therefore, the distributions of completion time and score for the two test media will be comparable for all three subject areas if the English score distribution for the CBT is equated to that for the PPT.

We hope to examine the causes of the significant effects of the test media found only on the scores for the English tests in the future. For English

1) the score distribution for the CBT was higher than that for the PPT

2) ANOVA revealed the significant main effect of the test media on score (see Table 3), and

3) the mean of score of the CBT was significantly higher than of the PPT in the PPT-first group (Table 5).

In the results of this study, no significant difference was found for the score of Japanese tests. The significant main effect of test media on the score

for Japanese tests in the previous study (Fujiyoshi & Ishizuka, 1996) seems to be caused by the problems of the previous experimental design which is similar to the design I (Angoff, 1971). Thus the significant main effect was not necessarily related to the differences of the two test media, such as differences of legibility for the test booklet and the computer screen, as well as the ease of turning over pages for the PPTs and the CBTs.

We found a general tendency on the effects of order of administration of the tests. The means of the completion time for the second tests were smaller than those for the first tests (see Table 4), and the means of the score for the second tests were slightly lower than those for the first tests except for English (see Table 5).

The mean of completion time of the second test for the total group for Japanese was significantly smaller than of the first test (see Table 4). This is believed to be due to the behavior of subjects in the exceptional situation of the experiment. As the Japanese tests were administered at very end of the experiment session, subjects who completed the test early for Japanese were able to leave the room and return home immediately. Therefore some subjects obviously tended to leave early and it is thought that this affected the completion time of the test.

The item cumulative time-completion rate curves and item cumulative time-score rate curves were newly developed for this research as a means of directly comparing the distributions of completion time and score for the PPTs and the CBTs graphically (see Appendix). The item cumulative time-completion rate curves were created from the distribution of the single factor of completion time. On the other hand, the item cumulative time-score rate curves were created from distributions of the two factors of completion time and score. The item cumulative time-completion rate curves and the item cumulative time-score rate curves are notably smooth and stable, even with small numbers of subjects (see Figure 4 and Figure 5). The numbers of points used in plotting the cumulative item time-completion rate curves are a few hundred times greater than the numbers of points required for the group learning response curves (Fujita, 1975; Fujiyoshi, 1997, 2000) (see Table 6), and the numbers of points on the item cumulative time-score rate curves are a few ten times greater than the numbers of points required for the time-score rate curves (Fujiyoshi, 1999, 2000) (see Table 7). The optimal distribution function fitted to these curves is currently under consideration. The Weibull distribution function does not always fit to these curves (Fujiyoshi, 2000), though it has been appropriate for the group learning response curves and time-score rate curves (Fujita, 1975; Fujiyoshi, 1997, 1999, 2000; Fujiyoshi & Ishizuka, 1996).

4 Conclusion

It was concluded from the results of this study that the distributions of completion time and score for the PPT and the CBT generally approximated each other for the Japanese, Mathematics and English tests in the National Center Test for University Admissions. The distributions for both test media will be comparable if the English score distribution for the CBT is equated to that for the PPT. As Bunderson et al. pointed out (Bunderson, Inouye & Olsen, 1989), linear CBTs were intended to serve the same purposes as their PPTs. Therefore, a key issue was the comparability of scores obtained on CBTs and PPTs of the same tests. Can the scores from the two test media be used interchangeably to make academic decisions? In the result of this study, except for the English score distributions, the results of ANOVA and t-test showed that there were no significant differences in the amounts of completion time and test scores for both test media for all three subject areas.

This assessment marked the first use of the notions of the item cumulative time-completion rate curves and the item cumulative time-score rate curves to compare the distributions for both test media. The shapes of curves for the two test media

were found to approximate each other.

The use of this CBT opens up possibilities of research into the answering process of PPTs. The conventional video recording method requires considerably much time and effort to collect the test data of the answering process, and has proved to be an obstacle to research into the answering process of PPTs. For the purpose, employing a pen computer, the CBT system was developed and designed to simulate conventional PPTs as faithfully as possible. The use of the CBTs allowed automated collection of all test data, and enabled to estimate the answering process of PPTs from test data collected by the CBTs.

The development of this CBT resolved the problems of the user interface in previous CBTs (Lee, 1986; Mazzeo et al, 1992). The use of the pen computers allows the user to add handwritten notes to the computer screen while answering questions, and to touch the marking sheet field on the screen with the electronic pen to answer questions directly. In comparison to indirectly operated pointing devices such as a mouse and a touch-pad, the pen computer considerably enhances simplicity of operations. The computer version of TOEFL (Test of English for Foreign Language) has employed a mouse and a keyboard and required the training of computer tutorial before testing (Eignor, Tailor, Kirsch & Jamieson, 1998; Kirsch, Jamieson, Tailor & Eignor, 1998; Tailor, Jamieson, Eignor & Kirsch, 1998). The results of the survey of questionnaire about this CBT system showed that the students who were administered the test (i.e. first year university students) accepted the CBTs positively. The ease of reviewing the page on the screen and of using the electronic pen were evaluated as being similar to that for the conventional PPTs employing marking sheets (Fujiyoshi, 2000; Fujiyoshi & Ishizuka, 1996).

The comparability of distributions of completion time and score for the two test media is not a simple matter of comparing distributions in terms of means, dispersions and shapes; it is also necessary to compare the rank orders of subjects according to the Guidelines for CBT and Interpretations of American Psychological Association (Mazzeo et al, 1992). The use of the Latin-square design in this experimental assessment precluded a comparison of the rank orders of subjects, and it is hoped that an investigation of equating by information on the rank orders of subjects will be conducted in the future.

Appendix. The Definitions of Item Cumulative Time-Completion Rate Curves and Item Cumulative Time-Score Rate Curves

Item cumulative time-completion rate curves and item cumulative time-score rate curves are defined as follows.

Suppose that there are n subjects ($subject_1$, $subject_2$,..., $subject_n$) and m items ($item_1$, $item_2$,..., $item_m$). For each $item_1$, $item_2$,..., $item_m$, suppose also that the score of $item_j$ is $score_j$.

Definition A.1 Let $t_{i,j}$ be the time that $subject_i$ completed $item_j$. If $subject_i$ didn't answer $item_j$, let $t_{i,j} = \infty$. Let $t_{\max}$ be $\max\{ t_{i,j} \mid 1 \leqq i \leqq n, 1 \leqq j \leqq m, t_{i,j} \neq \infty \}$. Let $c_{i,j}$ be an integer such that if $subject_i$ answered $item_j$ correctly, $c_{i,j} = 1$, otherwise $c_{i,j} = 0$.

Definition A.2 The item completion function for $subject_i$ and $item_j$, $\mathrm{IC}_{i,j}$ is a function such that if $t < t_{i,j}$, then $\mathrm{IC}_{i,j}(t) = 0$, else if $t_{i,j} \leqq t$, then $\mathrm{IC}_{i,j}(t) = 1$.

Definition A.3 The item cumulative time-completion function for $subject_i$, ICTC_i is defined as follows.

$$\mathrm{ICTC}_i(t) = \sum_{j=1}^{m} \mathrm{IC}_{i,j}(t)$$

The item cumulative time-score function for $\mathrm{subject}_i$, ICTS_i is defined as follows.

$$\mathrm{ICTS}_i(t) = \sum_{j=1}^{m} c_{i,j} \cdot score_j \cdot \mathrm{IC}_{i,j}(t)$$

Definition A.4 The item cumulative time-completion rate function ICTCR is defined as follows.

$$\mathrm{ICTCR}(t) = \sum_{i=1}^{n} \mathrm{ICTC}_i(t) \Big/ \sum_{i=1}^{n} \mathrm{ICTC}_i(t_{\max})$$

The item cumulative time-score rate function ICTSR is defined as follows.

$$\mathrm{ICTSR}(t) = \sum_{i=1}^{n} \mathrm{ICTS}_i(t) \Big/ \sum_{i=1}^{n} \mathrm{ICTS}_i(t_{\max})$$

Definition A.5 An item cumulative time-completion rate curve is a set of points on coordinate system with time t on the horizontal axis and value of item cumulative time-completion rate ICTCR(t) on the vertical axis. An item cumulative time-score rate curve is a set of points on coordinate system with time t on the horizontal axis and value of item cumulative time-score rate ICTSR(t) on the vertical axis.

Reference

Angoff, W. H. 1971 Scales, Norms, and Equivalent Scores. In Thorndike, R. L. (Ed.), *Educational Measurement.* 2nd ed. Washington, D. C.: American Council on Education. Pp. 508-600.

Bunderson, V. C., Inouye, D. K., & Olsen, J. B. 1989 The Four Generations of Computerized Educational Measurement. In Linn, R. L. (Ed.), *Educational Measurement.* 3rd ed. New York: Macmillan. Pp. 367-407.

Eignor, D., Tailor, C., Kirsch, I., & Jamieson, J. 1998 *Development of a Scale for Assessing the Level of Computer Familiarity of TOEFL Examinees. Educational Testing Service Research Report RR-98-7.* Princeton, NJ: Educational Testing Service.

Fujita, K. 1975 *An Introduction to Educational Information Technology.* Shokodo. (In Japanese)

Fujiyoshi, M. 1997 A New Method for Estimating the Time to Be Extended for Testing Students with Visual Disabilities from the Response Time Curves. *Research Bulletin The National Center for University Entrance Examinations,* **27,** 1-18. (In Japanese with English summary)

Fujiyoshi, M., 1999 An Improvement of Method to Estimate the Amount of Testing Time Extended for Students with Disabilities by Means of Time-Score Rate Curves. *Research Bulletin The National Center for University Entrance Examinations,* **9**, 31-37. (In Japanese)

Fujiyoshi, M. 2000 *An Experimental Study on the Estimation of Extension Rates of Testing Time for Students with Disabilities: Development of New Methods to Estimate the Extension Rates of Testing Time Quantitatively by Analyzing Answer Processes of Tests (For Test-Takers with Visual Disabilities as a Model).* Doctoral Dissertation of Mental and Physical Defectology (University of Tsukuba) (unpublished). (In Japanese)

Fujiyoshi, M., & Ishizuka, T. 1996 Development of Computerized Test system to analyze answer processes of Testing. *Research Journal of University Entrance Examinations,* **6,** 16-24. (In Japanese)

Kirsch, I., Jamieson, J., Tailor, C., & Eignor, D. 1998 *Computer Familiarity Among TOEFL Examinees. Educational Testing Service Research Report RR-98-6.* Princeton, NJ: Educational Testing Service.

Lee, J. 1986 The Effects of Past Computer Experience on Computerized Aptitude Test Performance. *Educational and Psychological Measurement.* **46**, 727-734.

Mazzeo, J., Druesne, B., Raffeld, P. C., Checketts, K. T., & Muhlstein, A. 1992 *Comparability of Computer and Paper-and-Pencil Scores for Two CLEPO General Examinations. Educational Testing Service Research Report RR-92-14.* Princeton, NJ: Educational Testing Service.

Tailor, C., Jamieson, J., Eignor, D., Kirsch, I. 1998 *The Relationship Between Computer Familiarity and Performance on Computer-based TOEFL Test Tasks. Educational Testing Service Research Report RR-98-8.* Princeton, NJ: Educational Testing Service.