Comparability of Paper-and-Pencil and Computer-Based Tests for Distributions of Completion Time and Score of the National Center Test for University Admissions

Mamoru Fujiyoshi*, Akio Fujiyoshi** and Tomoichi Ishizuka*

Abstract This experimental study was conducted to assess the comparability of conventional paper-and-pencil tests and the computer-based tests 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 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. The computer-based test employed a pen computer on which the paper-and-pencil test sheet was displayed on the computer screen, 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 administered the paper-and-pencil and computer-based tests differed, 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 distributions of completion time and score for the paper-and-pencil tests and the computer-based tests were comparable.

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

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, the pencil being replaced by the electronic pen.

The results of the experimental assessment in 1996 suggested that we should improve the experimental methodology. Distributions of completion time and score for Japanese, Mathematics, and English tests in the National Center Test for University Admissions differed minimally between the PPTs and the CBTs, however a significant difference was noted in the

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

score distributions for Japanese (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. The notions of item cumulative time-completion rate curves and item cumulative time-score rate curves were newly prepared in this study in order to compare the distributions of completion time and score for the PPTs and the CBTs. The results suggested that distributions of completion time and score were approximately the same for the PPTs and the CBTs, and that both were comparable.

2. The CBT

The CBT was designed to simulate the PPTs employing the optical readable marking sheets as used in the National Center tests as faithfully as possible. Pen computers (Amity SV) supplied by Mitsubishi Electric Corporation were employed in the development. 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. A handwritten memorandum has been added to 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, any handwritten memorandum added on the screen, and answers entered in the electronic marking sheet field on the screen, all with the electronic pen. Questions may be reviewed as often as necessary, and answers 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.

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

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

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. As 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 Entrance in Japanese, Mathematics and English.

The test sets used in the experiment consisted of set A and set B for each of Japanese, Mathematics and English, and were prepared from questions previously used in the National Center tests. The amount of questions in each test set was approximately equivalent to the amount of 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 in accordance with the work-limit method without time limit.

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 the completion time and scores. Results of ANOVA for completion time of the test are shown in Table 2. Results of ANOVA for test scores are shown in Table 3.

The distributions of the 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 test are shown in Figure 2. The '+' symbols shown in the box-and-whisker plots represent the mean of a distribution.

As shown in Table 2, the results of ANOVA for the completion time indicate that the main effects of the test media factor on the completion time were not significant for all three subject areas. The main effects of the subject group factor were

not significant. 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 main effect of the test media factor on the scores was not significant for Japanese and Mathematics but 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 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, administered the first test as a CBT, then administered the second test as a PPT. The remaining 16 subjects were assigned to the PPT-first group, administered the first test as a PPT, then 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 of 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 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 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 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 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 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. The item cumulative time-completion rate curves for the two test media for Japanese, Mathematics and English are shown in Figure 4. Item cumulative time-completion rate curve is a set of points on 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. Bold lines relate to PPTs, and 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.

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

3.3.4 Comparisons Using Item Cumulative Time-score Rate Curves

Item cumulative time-score rate curves were newly prepared to provide a more detailed comparison of score distributions for the PPTs and the CBTs. The item cumulative time-score rate curves for the two test media for the three subject areas are shown in Figure 5. 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. Bold lines relate to the PPTs, and thin lines relate to the CBTs.

The item cumulative time-score rate curves for the PPTs and the CBTs are very well matched each other for Japanese, Mathematics and English. The two curves for Japanese approximate, as are the two curves for Mathematics. Although a significant difference was apparent between the both means of score for English, as the score rate is used the two curves 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 of points for a subject on both curves.

3.4 Discussion

It was found in the results of this experimental assessment that the distributions of completion time and score of the PPTs and the CBTs were comparable for Japanese, Mathematics and English tests in the National Center Test for University Admissions, and that the significant main effect of test media on the scores for Japanese in the previous assessment (Fujiyoshi & Ishizuka, 1996) was not necessarily related to the difference of the two test media. The effects of the two test media on the completion time and scores were analyzed by ANOVA. Except for the scores for English, no significant main effects due to test media were apparent on the completion time and scores for the three subject areas (see Table 2 and Table 3). The item cumulative time-completion rate curves for the PPTs and the CBTs approximated each other closely 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).

The effects of order of administration of the tests on distributions of completion time and score were generally such that the means of completion time for the second tests were smaller than those for the first tests (see Table 4), and that the means of 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 considered to be due to the behavior of subjects in the exceptional situation of the experiment. As the Japanese tests were administered at the very end of the experiment session, subjects completing test for Japanese early were able to leave their seats 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 of 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 in detail. The item cumulative time-completion rate curves were prepared from the distribution of the single factor of completion time. On the other hand, the item cumulative time-score rate curves were prepared from distributions of the two factors of completion time and score. These curves were prepared from the cumulative values of number of items and scores each time any subject in any group answered an item, or answered an item correctly. The increase in relative cumulative frequency for items and the increase in relative cumulative scores for items answered correctly as time passed indicate a true representation.

Subject cumulative curves can be created from the item cumulative curves. If the amounts of completion time for any completion ratio are calculated for each subject in a subject group from the test data, subject cumulative time-completion rate curves for any completion ratio can be prepared from the collected data. Similarly, if the amounts of time for obtaining this score for any obtained score ratio are calculated for each subject in a subject group from the test data, subject cumulative time-score rate curves for any obtained score ratio can be prepared from the collected data. Subject cumulative time-completion rate curves for 100% completion ratio coincide with conventional group learning response curves (Fujita, 1975; Fujiyoshi, 1997; Fujiyoshi, 2000). The time-score rate curves (Fujiyoshi, 1999; Fujiyoshi, 2000) can be seen as subject cumulative time-score rate curves for 100% obtained score ratio. In addition to 100% completion ratio and 100% obtained score ratio, subject cumulative time-completion rate curves for any completion ratio as well as subject cumulative time-score rate curves for any obtained score ratio may also be considered. Further research is required on these characteristics.

The item cumulative time-completion rate curves and the item cumulative time-score rate curves are much 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 subject cumulative time-completion rate curves (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 subject cumulative time-score rate curves (see Table 7).

The optimal distribution function fitted to these curves is currently under consideration. The

Weibull distribution function is not always fitted to these curves (Fujiyoshi, 2000), though it has been appropriate for subject cumulative time-completion rate curves and subject cumulative time-score rate curves (Fujita, 1975; Fujiyoshi, 1997; Fujiyoshi, 1999; Fujiyoshi, 2000; Fujiyoshi & Ishizuka, 1996).

The Latin-square design is considered to be one of the more precise experimental designs to detect effects of test media. This assessment employed the Latin-square design to allow each subject to be administered the PPTs and the CBTs simultaneously in order to figure out the problems in the previous assessment. 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 Angofff (1971). In general, for a given sample size, greater precision is obtained from this design than from random groups or anchor-test designs. The Latin square design is almost similar to the single-group counterbalanced equating design. Therefore, the precision of Latin square design may be equal to or greater than of the single-group counterbalanced equating design.

It is hoped to examine the causes of the significant effects of the test media found only on the score distributions for the English tests. For English the score distribution for the CBT was higher than that for the PPT, and ANOVA revealed the significant main effect of the test media on the score distributions (see Table 3). Despite the general tendency, for English the mean of score of the later CBT was significantly higher than of the initial PPT in the PPT-first group as shown by the results of t-test (Table 5).

4. Conclusions

The study revealed that the distributions of completion time and score for the PPTs and the CBTs approximated each other in the experiment which used the Japanese, Mathematics and English tests in the National Center Test for University Admissions with marking sheets. It was concluded that the distributions of completion time and score for both test media were comparable. As Bunderson et al. (Bunderson, Inouye & Olsen, 1989) pointed out, 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? As a result of this assessment, except for the English score distributions, the results of ANOVA and t-test showed that there were no significant differences in distributions of completion time and score for both test media for all three subject areas.

This assessment was the first use of the notions of item cumulative time-completion rate curves and item cumulative time-score rate curves to compare the distributions for both test media, and 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 answering process of PPTs. The

conventional video recording method requires considerably much time and effort to collect test data of answering process, and has proved an obstacle to research into 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 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. The use of the pen computers allows the user to add a hand written memorandum 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 results of a survey of questionnaire also showed that the students administered the test (i.e. first year university students) positively accepted the CBTs. 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 user interface of this CBT system can be used in future for the National Center Test without training by means of any computer tutorials, though the computer version of TOEFL (Test of English for Foreign Language) has been required the training of computer tutorial before testing (Eignor, Tailor, Kirsch, & Jamieson, 1998; Kirsch, Jamieson, Tailor, & Eignor, 1998; Lee, 1986; Tailor, Jamieson, Eignor, Kirsch, 1998).

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, Druesne, Raffeld, Checketts, & Muhlstein, 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.

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 Examination,* **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 Examination,* **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.