인지과학, 제20권 제3호
Korean Journal of Cognitive Science
2009, Vol. 20, No. 3, 275~289.

A Computerized Testing system that Reduces Backward Reasoning in Multiple-choice Items*

Jooyong Park†

Department of Education Sejong University

A new computerized testing system, called the Computerized Multiple-choice Testing (CMMT) system, was introduced. In this system, questions of multiple choice (MC) items are presented first without options, so that students must generate answers for themselves. They can click for the options when they are ready, and can respond within a brief, specified time period. The present study was performed to examine whether this system is effective in reducing backward reasoning, I. e., using the options of MC items as cues to find the correct answer. One hundred and seventy-seven 6th grade students (12 year olds) were divided into two groups so that mean scores from a prior test were equal: The experimental group took an intervening computerized test in the new format, and the control group in the MC format. Five days after the computerized intervening test, a short answer paper-and-pencil final test was given. Testing effect was greater in the new system than in the MC system. Analysis of the final test response in relation to the intervening test response showed that i) the students retained the correct answer in the new system more than in the MC testing system, and that ii) students corrected their previous failures in the intervening CMMT format more than those in the MC format. These results suggest that the new system is effective in reducing backward reasoning.

Key words : Backward Reasoning, Computerized Testing, Elementary School Education

---

* This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund) (KRF-2007-B00130).

† 교신저자: 박주용, Department of Education Sejong University
E-mail: jpark@sejong.ac.kr

Introduction

With the advent of personal computers, technology usage in testing has continued to improve. Computerized testing has several advantages over paper-and-pencil testing. First, it is efficient: Grading and feedback can be given immediately after the test so that appropriate changes in teachers' instruction and/or students' studying strategies can be made promptly. Second, it can accommodate innovative methods: Computerization opens up unique possibilities that are not available on a paper-and-pencil test [1, 2]. It is possible, for example, to include non-text media such as graphics, audio, video, and animation. Moreover, interactivity, which refers to the reaction of the computer contingent upon human responses, can be used in a variety of ways. For instance, inserting a sentence in a paragraph utilizes interactivity. When a student selects a location, the paragraph is re-configured with the sentence. Computerized Adaptive Testing (CAT) is another example of an interactive testing method. The CAT system interacts with the student at the between-items level. Performance on a problem affects the difficulty level of the following problem. Another example of innovative testing technology can be seen in the automated essay grading systems, such as Latent Semantic Analysis (LSA) [3], and Bayesian Nets [4]. Some of these have already been commercialized (e.g., Criterion^{TM}, Intelligent Essay Assessor^{TM}).

Despite the recent explosion in research on computerized testing, most of the above-mentioned innovative testing techniques are not easily accessible to most teachers and test administrators. Therefore, many popular e-testing systems adopt the traditional MC format. The reason for the dominance of the MC format is its easy and objective scoring. In case of the short answer format, scoring is labor-intensive because semantic equivalence and partial credit have to be reckoned, requiring incredible time and energy.

However, there are some well-known limitations of the MC format. One of them is

guessing. In typical MC tests, examinees can use the options as cues to find the correct answer. By carefully reading the options, examinee can retrieve relevant knowledge, and eliminate implausible options to narrow down the correct answer. In addition to this, presenting options along with the stem has been criticized for its artificiality. That is, the setting is detached from everyday life. One clear example can be found in the title of a research paper on assessing physicians' competence, "Patients don't present with five options" [5]. These shortcomings of the MC format are brought about by the backward reasoning from the options to the answer.

To prevent backward reasoning, grid items in SAT math [6] and extended matching items have been invented. The special feature of the grid items is in the answer sheet. Instead of 4 or 5 slots in a row, each item is assigned some columns of 10 or so rows. In the case of 3 columns of 10 rows, when a student gets 213, he or she fills in 2 in the first column, 1 in the second column, and then 3in the third column. Some mathematical signs can also be included in the answer sheet ('/' to represent fractions as in 1/5). One problem is that it is suitable for numerical answers, but not for words.

Some words, such as medical terminology, can be easily handled in extended matching items [5]. The characteristic feature of extended matching items is in the number of its numerous options. Examinees are asked to locate their response on a booklet that contains thousands of options listed alphabetically, along with unique numbers (e.g., heart #3537). Examinees enter the number in their answer sheet. The main problem of this method is that lack of flexibility to deal with other words, except for terminologies.

Another option to deal with drawbacks of the MC format is to use innovative computerized technology. C-rater can score short answer items automatically [7]. It is used in the National Assessment for Educational Progress (NAEP) and a state-wide assessment in Indiana. However, in order to score short answer items automatically, a

lot of extra work by content experts is needed. In the case of c-rater, for each item a group of expressions with equivalent meaning first have to be identified. Another simple solution to the backward reasoning is to use a newly developed computerized testing system called the Computerized Modified Multiple Choice (CMMT) system [8].

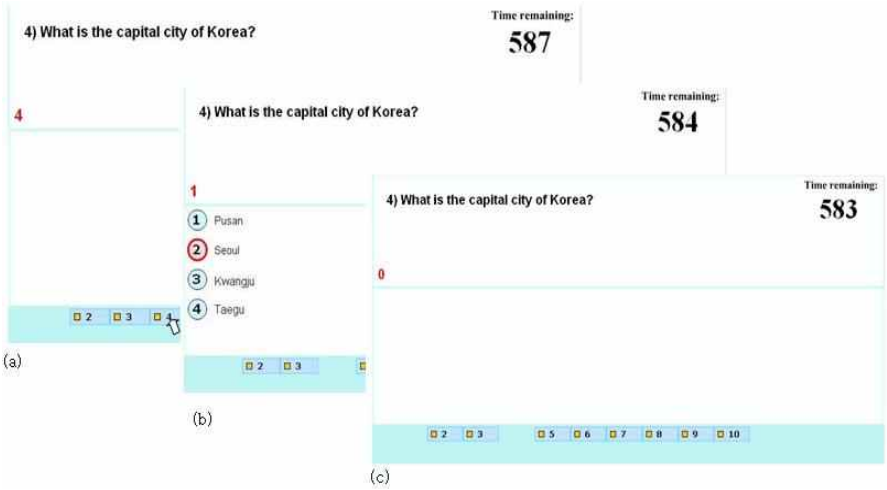

Figure 1. A sample demonstration of the computerized modified multiple-choice test. (a) Because question 1 has been solved, Box 1 has disappeared. The mouse is on Box 4, and Question 4 ("What is the capital city of Korea?") appears on the screen. The number 4 at the bottom of the question display is the preset time for responding to Question 4. (b) When the student clicks the mouse, the options are presented, and the time counts down to 0. During that time the student can make a selection and change it as many times as time allows. The selected option is marked. (c) When the time is over, the options disappear, and the student can no longer respond. (*Note.* From "Higher retention after a new take-home computerised test," by J. Park & B. Choi, 2008, *British Journal of Educational Technology, 39*, p.540. Copyright 2007by British Communications and Technology Agency. Reprinted with permission.).

The CMMT system was proposed to force examinees to generate their answer as if they were solving a short answer item. The system employs the interactivity of the

computer to decouple the tight connection between the stem and the options of the MC format. Once the examination starts, the student can see the questions without the response options (Figure 1a). When the student is ready to answer a question, the student can ask the computer to show its options by clicking the mouse button. The options appear on the screen for a pre-set amount of time. The pre-set time is shown on the screen, so that the student knows how long it is. The student can respond to the problem within that pre-set time (Figure 1b). The pre-set time is long enough for the student to check his or her answer against the options and choose the right option. When the pre-set time is up, both the question and the options disappear (Figure 1c). The crux of this manipulation is to force students to generate their own answers to the problems prior to recognizing their answer from the MC options. This procedure, as a whole, allows for active thinking while at the same time permits objective scoring [8, 9].

In previous studies, the usefulness of the CMMT as a learning aid has been shown by the enhanced testing effect [8, 9]. The testing effect refers to the enhanced retention of the learned material due to intervening tests [10. 11]. The testing effect is affected by test formats: It is greater for short answer or recall tests than for MC tests effect [10, 12]. Glover proposed a hypothesis to explain this difference. According to the hypothesis, the greater the retrieval load for the intervening test, the greater the testing effect.

Based on this hypothesis, Park (2005) predicted that the CMMT enhances memory retention more than the traditional MC test, because the CMMT system would activate a recall process at least before the options are presented. He found that the recall final test performance was, in fact, better after an intervening CMMT than that on an intervening MC test. This result has been replicated in a different environment by Park and Choi (2008). When the students took the intervening tests at home as a homework assignment instead of during a regular class, the CMMT format again

produced better final test scores than those on the traditional MC format.

The goal of this study is to test the hypothesis that the CMMT system is effective in reducing backward reasoning. As mentioned earlier, backward reasoning refers to guessing the correct answer by using the options presented along with the question (or stem) of the MC items. Because the options are presented briefly, it was expected that the possibility of students' using guessing strategies in the CMMT system is reduced. The methodology of comparing the CMMT and MC systems in this study is adapted from the research by Anderson and his colleagues who used conditional probabilities [13, 14]. Because the focus of Anderson and his colleagues was on a different issue, the logic of the current study is presented below without going into the details of their research.

The improved final test score as a result of testing effect can be represented as follows: P(C2) is greater in the CMMT than in the MC format, where C2 means the correct response to the final test. Now, the P(C2) can be partioned into the conditional probability of being correct on the final test given a correct answer on the intervening computerized test, P(C2|C1), and the conditional probability of being correct on the final test given a wrong answer on the intervening computerized test, P(C2|W1). The former conditional probability shows a better retention of the correct answer, and the latter probability implies correcting the previous failure. Because the options were presented briefly, it was hypothesized that guessing strategies such as using the correct option as a cue of the correct answer, and eliminating incorrect options, could be reduced in the CMMT system. If that were the case, we could predict that P(C2|C1) be higher in the CMMT system than that in the traditional MC testing system.

Method

Participants and Design

One hundred and seventy-seven students participated in the experiment. They were 6th graders attending J elementary school located in Seoul, Korea. They participated in the experiment as a part of their computer class activity to experience a new computerized testing method. These students had already taken both CMMT and MC format tests two times each. In these studies, six classes were divided into 2 groups so that there was no significant difference between the two groups in their mean scores of social studies at the end of the previous semester (69 vs. 71, n_s = 87 and 86 respectively). In this experiment, students who had transferred to the school during the 2nd semester were added, making the total number of the control group 88 and the experimental group 89.

The experimental group took the intervening test in the CMMT format, and the control group in the MC format. Students were familiar with the respective computerized testing system.

Apparatus

The software used for creating and running the experiment was Flash MX. The Flash files were displayed using the Microsoft Internet Explorer. The server and the databases were set up on a commercial web-hosting service. At the log-in page, students had to type in their name and password to log in to the system. The system did not allow a student to log in twice with the same name and password.

Material

The academic subject tested was 6thgrade social studies. There was one computerized

intervening test. The test had 20 questions and covered one unit that the students had learned during the previous two weeks. Most of the questions were on factual knowledge. The total testing time was 20 minutes, and the pre-set times for individual items were set between 4-7 secondsby the experimenter.

The same set of questions was given to the two different groups. The two main differences between the CMMT and the multiple-choice exams were i) whether the options were presented at the same time with the questions, and ii) whether revision was possible. In a CMMT item, the question alone was presented with a preset time in red; when the student wanted to respond to a question, he or she had to press the corresponding box below, thereby displaying the options, and respond to them within the preset time; when the time was over, the options disappeared and the student could no longer respond. The test score only appeared on the screen as feedback either when the student had answered all the questions or when the time was up. In the MC system, both question and its options were presented with a preset time in red, and examinees could revise their answer by reactivating the corresponding box below, whose color was marked red once they responded to the item. The feedback screen displayed the test score only. The final paper-and-pencil recall test had 10 questions, which were all odd numbered items on the intervening test.

Procedure

Students were trained to log in to the system with their own name and password and to answer the questions in their respective testing systems. The computerized test took place in the computer lab during the regular computer class under the supervision of the computer class instructor. The instructor did not know the intent of the present study, but learned how to administer the tests. Students were told to do their best on the test without any mention of its importance on their grades. Students were also

instructed to respond to the questions in any order they wanted. The final paper-and-pencil short answer test was given without notice 5 days after the computerized test. It was given in the morning before regular classes began, in each of the four classrooms simultaneously. Students were given 15 minutes for the final test.

Results

Among the 177 students, one student from the CMMT group did not finish the computerized test, and another 4 students (two from each group) did not take the paper-and-pencil final test. They were excluded from the analysis. Mean scores of the two groups with a total of 172 students are given in Table 1. The group which took the test in the traditional MC format received higher scores than the CMMT group in the intervening test ($t(170)=3.24$, $MSe=15.2$, $p<.01$, effect size=0.54). Also, when the average of the ten odd numbered items in the intervening test that were used in the final test were compared between the two groups, the MC group received higher scores than the CMMT group (29.3 vs. 25.8, $t(170)=2.4$, $MSe=9.5$, $p<.05$, effect size = 0.44).

However, there was little difference in the average proportion of correct responses

Table 1. Results of Experiment: Means and Standard Deviations of Test Scores (maximum 100 point for Total, and maximum 50 for the other two scores).

Types of Intervening tests	Test Scores		
	Int.test-Total	Int. test-Odd #s Only	Final test
Multiple-choice (n=86)	58.1 (13.9)	29.3 (8.3)	15.1 (9.6)
CMMT (n=86)	50.6 (16.4)	25.8 (10.6)	19.1 (9.8)

between the even numbered items and the odd numbered items (59 vs. 58; 52 vs. 50, MC and CMMT, respectively). Finally, the total testing time between the two groups were compared. The average testing time of the MC group was 494 seconds, which was significantly higher than the CMMT group, which was 396 seconds ($t(170)=3.98$, $MSe=160.4$, $p<.01$). This may be due to the fact that the examinee takes time to read the options and also uses time in selecting the right answer when an attractive distractor exists.

The CMMT group received significantly higher score on the test than the traditional testing group, $t(170)=2.6$, $MSe=15.1$, $p<.05$, (effect size = 0.39). The enhanced testing effect in the CMMT system was replicated even without item level feedback after the intervening test.[1)]

The crucial findings with respect to the current study involve two conditional probabilities: The conditional probability of being correct on the test given a correct answer on the computerized test, P(C2|C1), and the conditional probability of being correct on the test given a wrong answer on the computerized test, P(C2|W1). They are given in Table 2. Since three of the student got a perfect score on the intervening test, their P(C2|W1) could not be computed. They were excluded in the analysis. The P(C2|C1) was significantly higher in the CMMT system than in the traditional multiple-choice system, $t(170)=2.7$, $MSe=0.25$, $p<0.01$ (effect size = .44). The P(C2|W1) was also significantly higher in the CMMT system than in the traditional multiple-choice system, $t(167)=2.6$, $MSe=0.23$, $p<0.05$ (effect size = .39).

1) One reviewer pointed out that the present result can be explained by the transfer-appropriate-processing view [15]. This claim, however, actually supports the retrieval view in this case because it presupposes the similarity of the processing between the CMMT format and the SA format. Besides, Carpenter & DeLosh (2006) have shown that the testing effect cannot be accounted for in terms of the transfer appropriate processing view.

Table 2. Means and Standard Deviations of the Two Conditional Probabilities, P(C2|C1) and P(C2|W1)

Types of Intervening tests	P(C2|C1)	P(C2|W1)
Multiple-choice (n=86)	.31 (0.24)	.23 (0.25)a
CMMT(n=86)	.42 (0.26)	.34 (0.27)b

a. n=85

b. n=84

Discussion

The present study was designed to probe whether the CMMT system is more effective in reducing the use of options as a means of getting the right answer. A positive answer to this question was obtained through the analysis of two conditional probabilities. The conditional probability P(C2|C1) was higher in the CMMT format than in the MC format. This means that the CMMT system is superior to the traditional MC testing system in promoting the retention of the correct answer on the computerized intervening test. The reason for the higher retention of the correct answer in the CMMT format can be attributed to the reduced guessing in the format compared to the MC format. In other words, there is a higher chance that examinees who chose the right option in the CMMT system actually know the correct answer than those in the MC system. Also another evidence that the CMMT format reduces guessing using information provided by the options can be found in the analysis of the conditional probability P(C2|W1), that is, the probability of getting a correct answer on the final test after getting a wrong answer in the intervening test. It was again higher in the CMMT format. This result suggests that even in the CMMT tests, examineescan use the information provided by the options though they were presented only for a short time. In other words, because of the limited response time, when the

examinee could not find the answer he generated among the options he had to choose the best possible one among them in a short time which may well turn out to be wrong. However, he could probably remember the options after the test was over and determine a more valid answer. On the other hand, in the traditional MC test, the examinee already went through such process of using option information during the test and thus similar improvement in performance would not occur. This data implies that using information from the options to choose the correct answer, i.e., backward reasoning, occurs in MC tests and that CMMT tests can reduce it.

Therefore, the results from the conditional probability analyses suggest, in combination, that the CMMT system is more effective in reducing the use of guessing as a means of getting the right answer. They also suggest that the CMMT system is a more accurate testing tool in measuring the ability to generate answer than the MC system.

Although the present study has shown that the CMMT format is better than the MC format in reducing guessing, research on the CMMT format is still at an early stage. First of all, the findings in this study should be replicated across different disciplines and age groups. Also, beyond the retention of tested materials, there is more of a need to develop items that can access higher level thinking using this system. Concurrent validity with essays or other types of testing needs to be explored.

It should be noted that the CMMT format has its own weaknesses. One of them is that not all MC items can be presented in the CMMT format. Choosing the best option from among the given options is such an example. True/False items or Lickert scale items are another examples. We can keep these items as they are.

Another shortcoming is that, although the CMMT system is now used for research and actual assessment settings, additional measures must be taken for those individuals who have high test anxiety and/or who are slow in their responses. One idea is to give them a chance to write their own answers after the options disappear.

As more computers become available to students at school, and as more computers are wired with high-speed internet connections, empirical research on the CMMT format can be easily performed. Once the validity and effectivenss of the CMMT system is established through these studies, the system can be easily utilized in large-scale examinations. In the long run, it is hoped that the CMMT would serve as both an accurate assessment tool and an effective learning tool in a variety of settings.

References

[1] Parshall, C. G., Davey, T., & Pashley, P. J. (2000). Innovative item types for computerized testing. In W. J. van der Linden and C. A.W. Glas (eds.). *Computerized adaptive testing: Theory and Practice* (pp.129-148). Kluwer Academic Publishers.

[2] Sireci, S. G., & Zenisky, A. L. (2006). Innoviative item formats in computer-based testing: In pursuit of construct representation. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development,* Lawrence Earlbaum Associate: Hillsdale, NJ.

[3] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 25,* 259-284.

[4] Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning, and Assessment, 1*(2).

[5] Veloski, J. J., Rabinowitz, H. K., Robeson, M. R., & Young, P. R. (1999). Patients don't present with five choices: An alternative to multiple-choice tests in assessing physician's competence. *Academic Medicine, 74,* 539-546.

[6] Braswell, J., & Kupin, J. (1993). Item formats for assessment in mathematics. In R.E. Bennett, & W. C. Ward. (Eds.), *Construction versus Choice in cognitive Measurement.* New Jersey: Lawrence Erlbaum Associate.

[7] Leacock & Chodrow, (2003). C-rater: Automatied scoring of the short-answer questions. *Computers and the Humanities, 37,* 389-405.

[8] Park, J. (2005). Learning in a new computerized testing system. *Journal of Educational Psychology, 97*(3), 436-443.

[9] Park, J., & Choi, B. (2008). Higher retentention after a new take-home computerized test. *British Journal of Educational Technology, 39*(3), 538-547.

[10] Glover, J.A. (1989). The "testing" phenomonon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81(3)*, 392-399.

[11] Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Psychological Science in the Public Interest, 2*, 31-74.

[12] Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*, 268-276.

[13] Kulhavy, R. W., & Anderson, R. C. (1972). Delay-retention effect with multiple-choice tests. *Journal of Educational Psychology ,63*(5), 505-512.

[14] Surber, J. R., & Anderson, R. C. (1976). Delay-retention effect in natural classroom settings. *Journal of Educational Psychology, 67*(2), 170-172.

[15] Blaxton, T. (1989). Investigating dissociations among memory measures: Support for a transfer appropriate processing framework. *Journal of Experimental Psychology: Learning, Memory and Cogntion, 15,* 657-668.

1 차원고접수 : 2009. 7. 31
2 차원고접수 : 2009. 9. 18
최종게재승인 : 2009. 9. 24

요약

선다형 문항에서 역행추리를 줄이는 컴퓨터화 검사 방식

박 주 용

세종대학교 교육학과

변형 선다형 방식으로 명명된 새로운 컴퓨터화 검사가 소개되었다. 이 방식에서는 선다형 문항의 답지가 없이 먼저 질문만 제시되어 수험자가 스스로 답을 생각해내도록 한다. 수험자가 답을 할 준비가 되면 마우스를 클릭하여 답지 제시를 요청할 수 있는데, 답지는 미리 정해진 짧은 시간 동안만 제시되고 수험자는 그 시간 내에 답지 중 하나를 선택하여 반응한다. 본 연구에서는 이 시스템이 선다형의 약점으로 알려진 역행추리 즉, 답지를 이용하여 정답을 추측해내는 전략을 줄일 수 있는지를 알아보고자 하였다. 6학년 학생 177명을 이전 시험 결과에서의 평균에서 차이가 없게 두 집단으로 나눈 다음, 실험 집단은 새로운 방식으로 통제 집단은 선다형 방식으로 시험을 보도록 한 다음, 5일 후에 일부 문항을 단답형으로 다시 시험을 보게 하였다. 중간 본 시험에서의 반응을 고려하여 최종 검사 점수를 분석한 결과, i) 선다형보다 새로운 방식으로 보았을 때 정답을 더 잘 유지하였으며, ii) 선다형보다 새로운 방식으로 시험을 보았을 때 전에 틀렸던 문항에서 더 적게 틀림을 발견하였다. 이 결과는 새로운 컴퓨터화 검사 방식이 역행추리를 줄이는데 효과적임을 시사한다.

주제어 : 역행 추리, 검퓨터화 검사, 초등 교육